UQAC
Université du Québec
à Chicoutimi

THÈSE

PRÉSENTÉ A

L'UNIVERSITÉ DU QUEBEC A CHICOUTIMI

COMME EXIGENCE PARTIELLE

DU DOCTORAT SCIENCES ET TECHNOLOGIES DE L'INFORMATION

PAR

HANNECH AMEL

Système de recherche d'information étendue basé sur une projection multi-espaces

Juillet 2018

Résumé

Depuis son apparition au début des années 90, le World Wide Web (WWW ou Web) a offert un accès universel aux connaissances et le monde de l'information a été principalement témoin d'une grande révolution (la révolution numérique). Il est devenu rapidement très populaire, ce qui a fait de lui la plus grande et vaste base de données et de connaissances existante grâce à la quantité et la diversité des données qu'il contient. Cependant, l'augmentation et l'évolution considérables de ces données soulèvent d'importants problèmes pour les utilisateurs notamment pour l'accès aux documents les plus pertinents à leurs requêtes de recherche. Afin de faire face à cette explosion exponentielle du volume de données et faciliter leur accès par les utilisateurs, différents modèles sont proposés par les systèmes de recherche d'information (SRIs) pour la représentation et la recherche des documents web. Les SRIs traditionnels utilisent, pour indexer et récupérer ces documents, des mots-clés simples qui ne sont pas sémantiquement liés. Cela engendre des limites en termes de la pertinence et de la facilité d'exploration des résultats. Pour surmonter ces limites, les techniques existantes enrichissent les documents en intégrant des mots-clés externes provenant de différentes sources. Cependant, ces systèmes souffrent encore de limitations qui sont liées aux techniques d'exploitation de ces sources d'enrichissement. Lorsque les différentes sources sont utilisées de telle sorte qu'elles ne peuvent être distinguées par le système, cela limite la flexibilité des modèles d'exploration qui peuvent être appliqués aux résultats de recherche retournés par ce système. Les utilisateurs se sentent alors perdus devant ces résultats, et se retrouvent dans l'obligation de les filtrer manuellement pour sélectionner l'information pertinente. S'ils veulent aller plus loin, ils doivent reformuler et cibler encore plus leurs requêtes de recherche jusqu'à parvenir aux documents qui répondent le mieux à leurs attentes. De cette façon, même si les systèmes parviennent à retrouver davantage des résultats pertinents, leur présentation reste problématique.

Afin de cibler la recherche à des besoins d'information plus spécifiques de l'utilisateur et améliorer la pertinence et l'exploration de ses résultats de recherche, les SRIs avancés adoptent différentes techniques de personnalisation de données qui supposent que la recherche actuelle d'un utilisateur est directement liée à son profil et/ou à ses expériences de navigation/recherche antérieures. Cependant, cette hypothèse ne tient pas dans tous les cas, les besoins de l'utilisateur évoluent au fil du temps et peuvent s'éloigner de ses intérêts antérieurs stockés dans son profil. Dans d'autres cas, le profil de l'utilisateur peut être mal exploité pour extraire ou inférer ses nouveaux besoins en information. Ce problème est beaucoup plus accentué avec les requêtes ambigües. Lorsque plusieurs centres d'intérêt auxquels est liée une requête ambiguë sont identifiés dans le profil de l'utilisateur, le système se voit incapable de sélectionner les données pertinentes depuis ce profil pour répondre à la requête. Ceci a un impact direct sur la qualité des résultats fournis à cet utilisateur.

Afin de remédier à quelques-unes de ces limitations, nous nous sommes intéressés dans ce cadre de cette thèse de recherche au développement de techniques destinées principalement à l'amélioration de la pertinence des résultats des SRIs actuels et à faciliter l'exploration de grandes collections de documents. Pour ce faire, nous proposons une solution basée sur un nouveau concept d'indexation et de recherche d'information appelé la projection multi-espaces. Cette proposition repose sur l'exploitation de différentes catégories d'information sémantiques et sociales qui permettent d'enrichir l'univers de représentation des documents et des requêtes de recherche en plusieurs dimensions d'interprétations. L'originalité de cette représentation est de pouvoir distinguer entre les différentes interprétations utilisées pour la description et la recherche des documents. Ceci donne une meilleure visibilité sur les résultats retournés et aide à apporter une meilleure flexibilité de recherche et d'exploration, en donnant à l'utilisateur la possibilité de naviguer une ou plusieurs vues de données qui l'intéressent le plus. En outre, les univers multidimensionnels de représentation proposés pour la description des documents et l'interprétation des requêtes de recherche aident à améliorer la pertinence des résultats de l'utilisateur en offrant une diversité de recherche/exploration qui aide à répondre à ses différents besoins et à ceux des autres différents utilisateurs.

Cette étude exploite différents aspects liés à la recherche personnalisée et vise à résoudre les problèmes engendrés par l'évolution des besoins en information de l'utilisateur. Ainsi, lorsque le profil de cet utilisateur est utilisé par notre système, une technique est proposée et employée pour identifier les intérêts les plus représentatifs de ses besoins actuels dans son profil. Cette technique se base sur la combinaison de trois facteurs influents, notamment le facteur contextuel, fréquentiel et temporel des données.

La capacité des utilisateurs à interagir, à échanger des idées et d'opinions, et à former des réseaux sociaux sur le Web, a amené les systèmes à s'intéresser aux types d'interactions de ces utilisateurs, au niveau d'interaction entre eux ainsi qu'à leurs rôles sociaux dans le système. Ces informations sociales sont abordées et intégrées dans ce travail de recherche. L'impact et la manière de leur intégration dans le processus de RI sont étudiés pour améliorer la pertinence des résultats.

Mots clés: Recherche d'information, projection multi-espaces, facettes de données, Web sémantique, contexte de recherche, personnalisation de données, modélisation utilisateur, Web social, recommandation de données.

Abstract

Since its appearance in the early 90's, the World Wide Web (WWW or Web) has provided universal access to knowledge and the world of information has been primarily witness to a great revolution (the digital revolution). It quickly became very popular, making it the largest and most comprehensive database and knowledge base thanks to the amount and diversity of data it contains. However, the considerable increase and evolution of these data raises important problems for users, in particular for accessing the documents most relevant to their search queries. In order to cope with this exponential explosion of data volume and facilitate their access by users, various models are offered by information retrieval systems (IRS) for the representation and retrieval of web documents. Traditional SRIs use simple keywords that are not semantically linked to index and retrieve these documents. This creates limitations in terms of the relevance and ease of exploration of results. To overcome these limitations, existing techniques enrich documents by integrating external keywords from different sources. However, these systems still suffer from limitations that are related to the exploitation techniques of these sources of enrichment. When the different sources are used so that they cannot be distinguished by the system, this limits the flexibility of the exploration models that can be applied to the results returned by this system. Users then feel lost to these results, and find themselves forced to filter them manually to select the relevant information. If they want to go further, they must reformulate and target their search queries even more until they reach the documents that best meet their expectations. In this way, even if the systems manage to find more relevant results, their presentation remains problematic.

In order to target research to more user-specific information needs and improve the relevance and exploration of its research findings, advanced SRIs adopt different data personalization techniques that assume that current research of user is directly related to his profile and / or previous browsing / search experiences. However, this assumption does not hold in all cases, the needs of the user evolve over time and can move away from his previous interests stored in his profile. In other cases, the user's profile may be misused to extract or infer new information needs. This problem is much more accentuated with ambiguous queries. When multiple POIs linked to a search query are identified in the user's profile, the system is unable to select the relevant data from that profile to respond to that request. This has a direct impact on the quality of the results provided to this user.

In order to overcome some of these limitations, in this research thesis, we have been interested in the development of techniques aimed mainly at improving the relevance of the results of current SRIs and facilitating the exploration of major collections of documents. To do this, we propose a solution based on a new concept and model of indexing and information retrieval called multi-spaces projection. This proposal is based on the exploitation of different categories of semantic and social information that enrich the universe of document representation and search queries in several dimensions of interpretations. The originality of this representation is to be able to distinguish between the different interpretations used for the description and the search for documents. This gives a better visibility on the results returned and helps to provide a greater flexibility of search and exploration, giving the user the ability to navigate one or more views of data that interest him the most. In addition, the proposed multidimensional representation universes for document description and search query interpretation help to improve the relevance of the user's results by providing a diversity of research / exploration that helps meet his diverse needs and those of other different users.

This study exploits different aspects that are related to the personalized search and aims to solve the problems caused by the evolution of the information needs of the user. Thus, when the profile of this user is used by our system, a technique is proposed and used to identify the interests most representative of his current needs in his profile. This technique is based on the combination of three influential factors, including the contextual, frequency and temporal factor of the data.

The ability of users to interact, exchange ideas and opinions, and form social networks on the Web, has led systems to focus on the types of interactions these users have at the level of interaction between them as well as their social roles in the system. This social information is discussed and integrated into

this research work. The impact and how they are integrated into the IR process are studied to improve the relevance of the results.

Keywords: Information retrieval, multi-spaces projection, data facets, semantic web, search context, data customization, user modeling, social web, data recommendation.

Remerciement

Je tiens tout d'abord à remercier Dieu le tout puissant de m'avoir aidé et donné la force d'établir et concrétiser ce travail malgré les nombreux obstacles qu'il y avait eu.

Je profite de cette occasion pour présenter mes respects et mes remerciements à mes directeurs de thèse : Dr. Mehdi Adda et Dr. Hamid Mcheick pour m'avoir accueilli au sein de leurs groupes respectifs, pour leur encadrement et leurs précieux conseils tout au long de ces années de doctorat.

Je remercie monsieur Marc Gravel l'ancien responsable du programme de doctorat, d'avoir pris soin d'examiner ma candidature, et de m'avoir accordé la chance de réaliser ce que j'ai toujours rêvé de faire.

Je souhaite exprimer ma gratitude aux membres du jury pour avoir bien voulu accorder une partie de leur temps à lire ma thèse. Je remercie Mr. Sehl Mellouli, Mr. Mohamed Tarik Moutacalli et Mr. Abdenour Bouzouane d'avoir accepté de rapporter ma thèse. Toute ma gratitude s'adresse à Mr. Djamal Rebaine pour avoir accepté de présider le jury de cette thèse.

Je voudrais accorder une place d'honneur dans mes remerciements à ma famille, plus particulièrement, mes parents, et mes frères. Malgré la distance, leur amour et leur confiance me portent et me guident tous les jours, je ne serai pas là ou j'en suis sans eux.

Durant ces années de thèse, j'ai eu la chance de connaitre et de côtoyer de nombreuses personnes attachantes, je pense notamment à Fatima, Kenza, Asma, Ahlem, Imene et Moufida.

Je remercie ma chère amie Karima de m'avoir soutenu durant cette période de thèse, qu'elle trouve dans ces remerciements ma joie et mon bonheur de l'avoir dans ma vie. Je remercie Sara, une personne très spéciale à mon cœur, de m'avoir soutenu durant mes premiers années au canada, malgré que les conditions de la vie nous ont séparé aujourd'hui, qu'elle trouve dans ces remerciement mon plaisir et ma joie d'avoir partagé avec elle d'aussi agréables moments.

Enfin, le dernier et pas des moindres, je voudrai remercier particulièrement mon cher époux pour son amour, son soutien quotidien indéfectible et surtout sa patience à mon égard, je te remercie pour cette confiance, cette patience, merci d'être ce que tu es.

In the Name of ALLAH, the Beneficent, the Merciful

Surely my prayer and my sacrifice and my life and my death are surely for ALLAH, the lord of the words.

Le Sain Coran

Je dédie cette thèse à
mon mari, mes parents
et à mes deux frères

Table des matières

Liste des notations

RI	Recherche d'Information
RIW	Recherche d'Information sur le Web
SRIs	Systèmes de Recherche d'Information
RIF	Recherche d'Information par Facettes
RIP	Recherche d'Information Personnalisée
FC	Filtrage Collaboratif
CFS	Système de Filtrage Collaboratif
SRIP	Système de recherche d'Information Personnalisée
SRIF	Système de Recherche d'Information à Facettes
SFC	Sujets Fortement Connexes
RS	Réseau social

Liste des figures

Liste des tableaux

Chapitre 1 : Introduction générale

Ce travail s'inscrit dans le cadre de la recherche d'information sur le Web (RIW), en particulier la recherche d'information (RI) facettée, appelée aussi la RI par facettes (RIF), et vise à contribuer à l'amélioration de la qualité des résultats de recherche. Ceci est réalisé à travers la considération de plusieurs dimensions, appelées les vues de description et de recherche du contenu web, pouvant assister l'utilisateur dans ses quêtes d'information. Chaque dimension correspond à une façon de représentation différente de ce contenu, et permet de le découvrir et de l'explorer différemment. Les contributions de cette thèse se résument par les points suivants :

- La proposition d'une nouvelle approche de RI multidimensionnelle basée sur les facettes de données (cf. section 1.2.1 et section 1.3.1).
- La proposition d'un modèle utilisateur qui représente les centre d'intérêt de cet utilisateur qui sont appris derrière ses interactions avec le système de RI (cf. section 1.2.2 et section 1.3.2).
- Le passage d'un model uni–utilisateur à un modèle collaboratif qui aide à la formation de groupes d'intérêts derrière les activités similaires des utilisateurs et supporte la recherche d'information collaborative (cf. section 1.2.3).
- La proposition d'un modèle de RI personnalisée qui exploite le profil de l'utilisateur dans le but d'améliorer ses résultats de recherche (cf. section 1.2.4 et section 1.3.3).
- La proposition d'un modèle de résolution du problème de démarrage à froid (cf. section 1.1.6), qui aide à apprendre et construire rapidement le profil d'un nouvel utilisateur sur le système et permet de l'assister durant ses premières recherches du système (cf. chapitre 6).

Afin de bien situer cette étude, nous posons au cours de cette introduction la problématique quant aux techniques de représentation et de recherche d'information web. La problématique est affinée dans des points spécifiques « les questions de recherche » permettant d'expliquer le contexte de notre étude, et de ressortir les objectifs qui aident à bien mener ce travail. Pour répondre aux questions qui restent soulevées dans la littérature de ce cadre de recherche et mettre en œuvre les fondements de la nouvelle

méthodologie proposée, une démarche de travail nous est nécessaire. Elle permet de déterminer les structures et les techniques adoptées pour atteindre les objectifs ciblés.

I.1. Questions de recherche

La problématique de la RIW peut être liée à plusieurs et différents niveaux du processus de recherche. Les limitations existantes dans ce domaine ouvrent la voie à plusieurs projets de recherche. Notre étude vise à adresser les défis qui sont liés principalement aux modèles d'indexation et de recherche du contenu web, et aux modèles d'organisation des résultats sur l'interface. Ces modèles sont proposés pour réduire l'impact du problème de la surabondance d'information sur le web et permettre à l'utilisateur de n'avoir accès qu'à l'information pertinente. D'autres limitations secondaires sont également adressées dans cette étude, elles sont liées à i) la collecte des données d'intérêt des utilisateurs utilisées pour construire leurs profils en vue de personnaliser leurs résultats de recherche et améliorer davantage leur pertinence, ii) au modèle de représentation de ce profil et à la technique de son intégration dans le processus de RI. Ce travail de recherche aborde aussi les limitations qui sont liées à l'intégration du web social dans la RI.

I.1.1. Surcharge/surabondance d'information

Avec l'évolution considérable des technologies web, une grande quantité d'informations devient de plus en plus accessible. Les moteurs de RIW rapportent des milliards de pages dans leur index qui est en constante augmentation. Le nombre d'utilisateurs quant à lui est estimé à plusieurs centaines de milliers, si ce n'est de millions, et il ne cesse de s'accroitre. En outre, le web a évolué d'une plateforme où le rôle des utilisateurs se limitait essentiellement à la consommation du contenu, à une plateforme où ils sont au centre même de la production et du partage du contenu. Ces facteurs ont soulevé des défis majeurs pour la tache de la recherche efficace de l'information. Leur augmentation continuelle nécessite impérativement la conception et la mise en œuvre des outils efficaces, permettant de faciliter l'exploration de grandes collections de données et améliorer la pertinence des résultats. Cette pertinence est au centre des problématiques des techniques de la RI. Elle correspond à la capacité qu'un système puisse retourner les informations qui répondent le plus aux besoins des utilisateurs.

I.1.2. Changement et évolution du besoin informationnel de l'utilisateur

La RI est un domaine qui étudie la manière de répondre pertinemment à une requête utilisateur en retrouvant les informations qui correspondent aux mieux à ses attentes. Naturellement, les besoins en information varient d'un utilisateur à l'autre et peuvent également changer et/ou évoluer chez le même utilisateur d'un moment à l'autre en fonction de plusieurs paramètres regroupés sous le terme du « contexte ». Ce contexte peut inclure plusieurs facteurs, tels que la localisation géographique de l'utilisateur (Said *et al.* 2011) (Bouidghaghen et Tamine 2012), sa démographie (Weber et Castillo 2010), le contexte de la recherche, appelée aussi l'activité de recherche (Shen *et al.* 2005) (Tamine-Lechani *et al.* 2008) (Daoud *et al.* 2010a) (Asfari 2011), le temps (Zhao *et al.* 2006) (Boughareb et Farah 2013a) (Mezghani *et al.* 2014), l'événement pouvant influer la recherche (Boughareb et Farah 2012) (Boughareb et Farah 2013b), etc. L'utilisation de ce contexte dans le domaine de la RIW est une voie de recherche prometteuse. Son importance a été reconnue dans plusieurs travaux (Alonso *et al.* 2007) (Pasca 2008) (Diaz 2009) (Dinh et Tamine 2012). Ces informations contextuelles peuvent être recueillies implicitement sur l'activité de navigation de l'utilisateur dans le but de mieux identifier son besoin en information, ou fournies explicitement par l'utilisateur lui-même. Aussi, la manière utilisée pour exprimer un même besoin d'information peut différer d'un utilisateur à l'autre, ceci dépend de ses acquis et ses expériences de recherche, et aussi de ses connaissances dans le domaine de la recherche cible (Allan *et al.* 2003). Ainsi, afin de bien assister les utilisateurs dans leur recherche, les systèmes de recherche d'information (SRIs) doivent d'une part être capables d'identifier le besoin en information de l'utilisateur et de s'adapter à ses différents intérêts et préférences ainsi qu'à leur évolution dans le temps, et d'une autre part de repérer les requêtes équivalentes qui correspondent à un même besoin d'information. Celles-ci peuvent être exprimées différemment par le même utilisateur ou par différents utilisateurs. Cette adaptation reste un grand défi essentiel et ouvert pour les SRIs.

I.1.3. Représentation du contenu informationnel des documents

I.1.3.1. Rigidité des modèles de représentation et de recherche d'information monodimensionnelles

Les modèles de représentation des données jouent un rôle important dans le processus de la RI pertinente. Ils contribuent à la réduction de l'écart entre les termes de la requête utilisateur et ceux utilisés dans la représentation des documents web (Krovetz 1997). Les modèles sémantiques viennent remédier aux limitations qui sont soulevées dans la recherche classique. Cette recherche classique s'appuie uniquement sur les mots clés pour décrire et chercher l'information, et ne permet pas de lier les informations entre elles. Ceci engendre une dégradation de performance du système. Pour ce faire, le domaine sémantique prend en considération le sens du mot en vue d'extraire une représentation riche qui couvre mieux le besoin informationnel de l'utilisateur exprimé par une requête de recherche, et décrit mieux l'information recherchée par le système. Ceci prend en compte les liens sémantiques entre les mots et permet de rapprocher les informations entre elles lors du processus de recherche. Par exemple, pour un mot donné, il peut y avoir plusieurs sens. Cette notion est connue sous le nom de la polysémie. L'un des sens de ce mot peut être le même qu'un autre mot (synonymie). Il peut exister un mot plus générique appelé l'hyperonyme, ou un mot plus spécifique, l'hyponyme, ainsi que d'autres liens sémantiques liant les mots entre eux.

Ce domaine a connu un nouveau souffle grâce à l'émergence de nouvelles technologies du web sémantique, en particulier les standards de description et d'enrichissement du contenu. Ces standards ont été utilisés pour rendre les documents web aussi accessibles et exploitables que possible. Cet enrichissement est possible grâce à l'usage de données écrites dans un vocabulaire prédéfini décrivant une sémantique explicite. Nous citons par exemple les ontologies de domaines et les ontologies linguistiques, les thésaurus, les dictionnaires lexicaux, les taxonomies, etc. (Amar 2009). Ceci permet entre autres de faciliter l'interaction entre les utilisateurs et le système. Cependant, en dépit des nombreux avantages qui ont pu être apportés et offerts par le web sémantique et qui n'ont pas été tous cités dans cette section, celui-ci souffre de plusieurs limitations importantes. L'indexation sémantique souffre de la non-disponibilité d'un vocabulaire unique d'indexation (Ralalason 2010). Cette faiblesse est rencontrée par exemple avec l'utilisation des ontologies de domaines qui fournissent juste un vocabulaire spécifique à un seul domaine de connaissance. Ainsi, le mécanisme d'indexation proposé doit être en mesure de s'adapter aux différentes ontologies rencontrées. D'autres faiblesses peuvent être aussi soulignées, elles sont liées aux modèles de représentation qui organisent les données sous forme de taxonomie ou de hiérarchie de

données. Cette représentation n'offre qu'une seule et unique relation (est-un) liant les données entre elles. Ceci peut réduire la découverte d'informations pertinentes lors de la recherche/exploration du contenu système (Amar 2009). En outre, dans ce type d'organisation chaque donnée est associée à une seule et unique classe ou catégorie de description. Les données sont organisées du plus général au plus spécifique et peuvent être découvertes à travers un seul parcours qui peut être parfois difficile et long à atteindre, en particulier avec les grandes hiérarchies de navigation. De plus, dans une telle hiérarchie de description, une donnée spécifique est accessible uniquement en parcourant les données les plus générales qui lui sont associées. Ceci rend impossible la navigation entre des concepts spécifiques sans passer par des concepts généraux. Cette rigidité peut rendre la recherche lente, désagréable et inefficace (Zheng *et al.* 2013).

Un autre problème qui peut être également soulevé avec ce paradigme de navigation est vu lors de la gestion du contenu système. Par exemple l'ajout ou la modification des documents web peut requérir une mise à jour du modèle hiérarchique sur lequel se base la représentation de leur contenu, notamment l'ajout de nouveaux concepts qui nécessite un expert pour leur intégration au modèle de représentation (Marleau *et al.* 2008) .

I.1.3.2. Modèles multidimensionnels standards et non personnalisés

Dans le but d'améliorer et de faciliter l'exploration de grandes collections de données, le paradigme de la RI à facettes, appelé aussi la recherche facettée, fut son apparition. Ce paradigme vient atténuer les limites imposées par les classifications à structure mono hiérarchique en offrant une classification multidimensionnelle basée sur les facettes. Cette technique de classification permet de mieux appréhender et valoriser les informations, en intégrant plusieurs dimensions pour décrire l'information, tout en donnant à l'utilisateur la possibilité de naviguer selon les caractéristiques qui l'intéressent le plus, dans l'ordre qui lui convient, et en lui offrant aussi la possibilité de modifier dynamiquement les caractéristiques de sa recherche (Hildebrand *et al.* 2006) (Evéquoz *et al.* 2010). Par exemple, dans le cas d'une base de données de voitures, les facettes peuvent être « couleur », « marque », « prix », etc. Chacune de ces facettes peut prendre une valeur « rouge », « bleu », etc. pour la facette « couleur », et une gamme « 0-1000 » « 1000-2000 », etc. pour la facette « prix ». Les méthodes de regroupement dynamiques des résultats de recherches sous des facettes sont une des techniques proposées dans cette direction de

recherche. Elles permettent aux utilisateurs de raffiner et d'explorer au mieux les résultats. Par exemple, dans le domaine des images architecturales, les résultats peuvent être regroupés par matériaux (béton, brique, bois, etc.), styles (baroque, gothique, etc.), types de vue, gens (artistes, architectes, etc.), les emplacements, les périodes, etc.

Malgré ces multiples qualités prometteuses, la recherche facettée est loin d'être suffisante, de nombreux problèmes découlent de son adoption (Fagan 2013). On distingue les limitations qui sont liées à l'interface utilisateur (Dugast 2011), notamment, la surcharge d'information causée par le retour d'un trop grand nombre de résultats sur l'interface (facettes et/ou valeurs de facettes). De plus, il est largement reconnu que l'utilisation d'un grand nombre de facettes sur l'interface peut déstabiliser l'utilisateur dans sa recherche et rend l'exploration ambigüe. C'est pourquoi l'organisation des données sur l'interface utilisateur représente un aspect essentiel pour aider l'utilisateur à explorer au mieux les réponses (Bonnel et Moreau 2005) et s'avère être une difficulté majeure pour les systèmes de la recherche facettée. Dans de tels cas, comment faire pour réduire le nombre de ces données surtout que les besoins des utilisateurs sont différents et une seule interface ne répond certainement pas à toutes leurs attentes ? Quelles sont les facettes pertinentes pour un utilisateur donné ? De nombreuses questions restent soulevées à ce sujet, elles sont liées à i) l'utilisabilité des interfaces de recherche proposées (cf. section I.1.8), et à ii) la pertinence des données (facettes et valeurs de facettes) proposées par ces systèmes sur l'interface utilisateur.

I.1.4. Pertinence du contenu pour un utilisateur

Différentes techniques ont été utilisées pour personnaliser et réduire les données sur l'interface utilisateur. Plusieurs approches intègrent les modèles utilisateurs connus aussi sous le nom des profils utilisateurs (Amato et Straccia 1999b) (Koren *et al.* 2008) (Sieg *et al.* 2007) (Challam *et al.* 2007) (Le *et al.* 2012) (Nguyen *et al.* 2016). Ces modèles représentent les caractéristiques spécifiques des utilisateurs qui peuvent être de différentes catégories (données personnelles, données d'activités, données géographiques, préférences, informations sociales, etc.). Chaque approche exprime et utilise différemment le modèle de l'utilisateur, et son intégration dans le processus de la RI s'effectue à plusieurs et différentes fins : réorganisation des facettes de données sur l'interface utilisateur, réordonnancement centré-utilisateur de la liste des résultats de recherche, filtrage de données, prédiction de données, etc.

Certaines d'autres approches ne se reposent pas sur le profil complet de l'utilisateur mais sur l'historique le plus récent de ses recherches (Fu et Kim 2013) (Mezghani *et al.* 2014).

Le principal problème qui se pose dans cette direction de recherche est lié aux capacités des approches proposées à estimer la pertinence d'un résultat pour un ou plusieurs utilisateurs. Cependant, une technique de personnalisation peut être mauvaise lorsque les données exploitées dans ce processus ne correspondent pas au contexte de la recherche courante de l'utilisateur exprimée par une requête de recherche. En effet, les attentes de l'utilisateur peuvent changer, et une nouvelle requête peut être liée à un nouveau besoin qui ne peut pas être déduit de son historique récent ou de son profil complet. Ces méthodes souffrent donc du changement de contexte. En fait, lorsqu'un utilisateur est à la recherche de quelque chose qui n'est pas lié à son historique de recherches antérieures, sa requête sera mal interprétée. Nous appelons cela par le problème du contexte de recherche déconnecté. Ce problème est beaucoup plus accentué avec les requêtes courtes ambigües, appelées aussi les requêtes polysémiques. Ces requêtes sont liées à plusieurs interprétations (ex. Java, avocat, virus, jaguar, win, etc.). Pour expliquer mieux ce qui vient d'être dit, imaginons qu'un utilisateur a déjà fait des recherches dans le passé sur les virus informatiques, lors de sa soumission d'une nouvelle requête sur les virus « types of virus », le SRI qui se base sur une telle personnalisation lui proposera des résultats en relation avec son profil qui portent sur les virus des ordinateurs en dépit des autres interprétations auxquelles est liée cette requête. Un tel cas est observé avec la liste des réponses provenant des moteurs de recherche populaires Google et Bing, lorsque l'utilisateur fait des recherches sur le langage de programmation Java, quand il tape à nouveau le mot « Java », les seules interprétations qui lui sont offertes dans les premières pages sont liées à la plate-forme java ou le langage de programmation. Cela rend la diversité sémantique presque inexistante. D'un autre côté, si un autre utilisateur fait des recherches à des instants différents sur respectivement le langage Java et le groupe de musiciens Java, lors de sa resoumission de la requête : « Java animation», sur quoi le système doit-il se baser pour inférer le besoin courant de cet utilisateur ? La plupart des moteurs de recherche privilégient les informations récentes de l'utilisateur pour déduire ses attentes, néanmoins cette démarche peut induire aussi en erreur.

En outre, les attentes de l'utilisateur évoluent et certaines données enregistrées dans son profil peuvent devenir obsolètes dans le temps (Zheng et Li 2011). Par exemple, une nouvelle maman peut être

intéressée par des produits dédiés aux bébés, au fil du temps les intérêts de cette maman vont progressivement changer vers d'autres intérêts. Ainsi, l'exploitation de ces données d'intérêt peut aussi induire à des résultats non pertinents.

La même chose peut être soulevée avec les systèmes de filtrage collaboratif qui se basent sur la détection des profils similaires pour personnaliser du contenu à l'utilisateur cible (Shardanand et Maes 1995) (Jiang *et al.* 2012) (Lee *et al.* 2012) (Sneha et Varma 2015). Les utilisateurs qui sont considérés similaires par le système peuvent être non pertinents pour une personnalisation pertinente lorsqu'ils ne représentent pas réellement la communauté d'intérêt de cet utilisateur cible. Cela peut être dû en raison de plusieurs facteurs. Par exemple, la difficulté à extraire le besoin courant de cet utilisateur, notamment avec les recherches ambigües, l'obsolescence des données dans les profils des utilisateurs, l'évolution du besoin en information de l'utilisateur, etc. Des limitations peuvent être aussi observées lors du processus du calcul de la similarité entre les utilisateurs lorsque le contexte de leurs recherches n'est pris en compte pour représenter leurs intérêts. Un tel cas se produit quand deux utilisateurs tapent les mêmes requêtes de recherche, mais explorent différents résultats liés à différentes interprétations, comme le cas des requêtes courtes et polysémiques, ou quand ils emploient les mêmes d'étiquettes pour annoter différentes ressources. Une simple comparaison entre ces utilisateurs, en termes de leurs historiques d'annotations ou de requêtes de recherche, considère les deux utilisateurs comme similaires, ce qui n'est pas réellement le cas. Ceci peut induire en erreur et rend la recherche collaborative inefficace. L'interprétation pertinente des intérêts des utilisateurs est donc primordiale afin de garantir une personnalisation adéquate. Une mauvaise interprétation influence négativement sur la construction de leurs profils, ceux-ci représentent l'élément principal de cette tache de personnalisation.

Comme nous pouvons le constater, l'utilisateur n'est pas toujours bien servi par ces systèmes de personnalisation. Ils ne sont pas toujours utiles pour résoudre le problème des différents contextes de recherche qui se produit suite à différentes situations parmi lesquelles celles que nous venons de soulever. Ainsi, afin d'offrir du contenu pertinent à l'utilisateur, une méthode de personnalisation doit être en mesure d'identifier le besoin en information courant de cet utilisateur et de bien interpréter et représenter ses différents intérêts dans son profil afin de bien correspondre ce besoin informationnel à ces intérêts stockés dans le profil. Elle doit être aussi en mesure de gérer automatiquement l'évolution et le

changement perpétuels de ces intérêts afin de garantir leur bonne exploitation. La négligence de ces facteurs engendre plusieurs limitations qui ont un impact direct sur la qualité des résultats fournis. L'atténuation de ces limitations représente l'un des défis de cette étude.

D'autres types de personnalisation sont les systèmes de recommandation qui tentent de prédire les intérêts des utilisateurs qu'ils n'ont pas encore considérés. Ils se basent alors sur les expériences de recherche de ces utilisateurs et/ou ceux des autres utilisateurs similaires de leurs communautés. Ces systèmes tentent de répondre aux questions suivantes : par quels domaines l'utilisateur pourrait-il s'intéresser ? Quels documents l'utilisateur veut-il découvrir ? Quels sont les documents que cet utilisateur n'a pas encore consultés et qui pourraient l'intéresser ? Et surtout quels sont les intérêts et les préférences pertinentes de cet utilisateur peuvent servir à déduire de nouveaux intérêts ? Plusieurs approches se basent pour la prédiction des données sur la construction des règles d'association (Adda *et al.* 2007) (Chandrakar et Saini 2015) (Beldjoudi *et al.* 2017), une voie sur laquelle nous nous basons dans nos propositions pour suggérer de nouveaux intérêts aux utilisateurs. Ces intérêts ne sont pas nécessairement liés à un besoin spécifique exprimé par l'utilisateur via une requête de recherche. Dans le domaine de la personnalisation web, l'extraction des règles d'association est une technique de la fouille de données la plus utilisée. Une règle d'association nous renseigne sur la dépendance entre les objets. Elle est de la forme $X \rightarrow Y$, où X et Y représentent un ensemble d'objets de contenu, qui peuvent être de différents types (documents visités, étiquettes utilisées (Beldjoudi *et al.* 2011b), produits achetés, domaines d'intérêt (Shaw *et al.* 2010), etc.), l'ensemble d'objets X est nommé la prémisse de la règle, et l'ensemble Y est la conclusion de la règle.

La complexité avec l'utilisation de telles règles est une des problématiques de notre recherche. Elle réside dans la difficulté de reconnaitre les règles pertinentes pour les utilisateurs. Ceci est dû au nombre important de règles qui peut être extrait depuis l'ensemble de données de départ, causé de son côté par le nombre important d'activités qui sont effectuées sur le web par les utilisateurs. Le développement de techniques de classification ou de sélection de ces règles s'impose dans un tel cas.

I.1.5. Collecte de données relatives à l'utilisateur

Comme expliqué plus haut, les données d'intérêt des utilisateurs peuvent être recueillies implicitement sur leurs activités de navigation (fichiers logs, par exemple), ou fournies explicitement par eux-mêmes à travers des formulaires, des questions, etc. (ex. le sexe, l'âge, lieu de résidence, domaines d'intérêt, etc.) (Pannu *et al.* 2013). Un des inconvénients des systèmes de personnalisation que nous pouvons alors rajouter dans cette section est lié à la saisie manuelle des profils utilisateurs qui représente un processus souvent long et ennuyeux pour ces utilisateurs. Une grande surcharge cognitive entraîne le plus souvent un abandon de leur part. En outre, la collecte implicite de données peut être difficile pour déceler les intérêts réels de ces utilisateurs. En général, ce processus se base sur l'observation et l'analyse du comportement des utilisateurs à travers le système pour construire les fichiers logs. Cette analyse est le plus souvent classique, elle est basée sur le nombre de clics utilisateurs effectués sur les documents web. Un document web est donc considéré comme pertinent par un utilisateur donné s'il a été fréquemment visité par lui. Cette hypothèse peut induire à l'erreur, car un utilisateur peut cliquer sur des documents sans avoir vraiment d'intérêt pour leur contenu (Mezghani *et al.* 2014). D'autre part, les systèmes utilisant les informations implicites, telles que les cotes, souffrent de manque de rétroactions de la part des utilisateurs, c'est à dire, la proportion des éléments évalués est souvent très faible.

Par ailleurs, le nombre important d'activités effectuées sur le web par les utilisateurs peut être aussi problématique et rend difficile la tâche de détection de leurs intérêts et préférences. Comment font donc ces systèmes pour faire face à toutes ces limitations ? Ces défis n'ont jamais été entièrement relevés et restent à ce jour soulevés (Jaseena et David 2014).

I.1.6. Démarrage à froid d'un nouvel utilisateur

Dans certaines situations, les intérêts de l'utilisateur sont indisponibles, en particulier, lorsque cet utilisateur utilise le système pour les premières fois. Ce dernier se retrouve incapable de lui suggérer du contenu personnalisé, et les réponses renvoyées peuvent être loin de ses attentes. Ce problème est connu sous le nom du démarrage à froid d'un nouvel utilisateur. Il est généralement traité en utilisant des sources de données externes capables d'alimenter le système, telles que les données sociodémographiques de l'utilisateur (Nguyen *et al.* 2006) (Meng *et al.* 2013) (Zhang *et al.* 2013), ses préférences qui sont

saisies manuellement (Stolze et Rjaibi 2001), etc. Comme nous pouvons le constater, ces méthodes nécessitent l'existence d'un minimum d'informations sur l'utilisateur qui doivent être éditées par lui-même ou détectées automatiquement lors de sa première connexion. Ces informations ne sont pas toujours disponibles pour des raisons de respect de la vie privée par exemple le cas des données sociodémographiques, ou tout simplement pour les raisons soulevées précédemment qui sont liées à la saisie manuelle du profil utilisateur. D'autres méthodes se basent sur les relations de confiance qui relient les utilisateurs les uns aux autres (Golbeck et Hendler 2006) (Haydar *et al.* 2012). Ainsi, même si l'utilisateur n'a toujours pas interagi avec le système, il pourra quand même recevoir des recommandations s'il est relié à d'autres utilisateurs. Évidemment, cela nécessite que cet utilisateur connaisse d'autres utilisateurs qui font partie de ce système, ce qui n'est malheureusement pas souvent le cas lorsque l'utilisateur est nouveau. Les performances de ces systèmes peuvent donc être très mauvaises en raison de l'absence de toutes ces informations sur l'utilisateur (son profil, son réseau social, etc.) sur lesquels se fondent généralement les systèmes de personnalisation pour proposer du contenu.

I.1.7. Impact de l'avènement du web interactif centré utilisateur : le web social

L'arrivée du web centré utilisateur, en particulier le web 2.0, tel que les blogs, les sites de partages, les réseaux sociaux, les wikis, etc., a changé le rôle des utilisateurs de consommateurs vers des producteurs de contenu. Ceci a fait émerger plusieurs pratiques dont l'étiquetage collaboratif qui peut constituer à la fois une aide pour les utilisateurs en les guidant vers l'information souhaitée, et pour le système afin de comprendre les intérêts de ces utilisateurs. Ce paradigme a été fortement développé ces dernières années, il représente la clé des pratiques sociales du Web 2.0 en permettant aux utilisateurs d'annoter en ligne les documents par l'usage de mots clés appelés les étiquettes ou les tags, et de construire ce qui est connu par les folksonomies. Ces étiquettes expriment les intérêts des utilisateurs, et aident entre autres à compléter les modes d'organisation et de navigation actuels qui sont souvent sous forme de cases à cocher, ou de menues déroulants, etc. Cependant, l'introduction de l'étiquetage pour chercher, explorer, ou encore inférer de l'information peut être à son tour une source de problèmes qui a un effet négatif sur le processus de découverte de relations entre les objets (Beldjoudi *et al.* 2011a), en particulier les relations qui relient les étiquettes aux documents, et celles qui existent entre les étiquettes.

Nous citons le problème de l'ambigüité des étiquettes qui se manifeste lorsqu'un terme a plusieurs interprétations (significations). Cela peut avoir un impact sur la pertinence des données découvertes à base de ce terme. Par exemple, lorsqu'un utilisateur utilise le terme « Java » pour annoter un document qui discute un tutoriel de programmation, une exploration basée sur cette étiquette peut avoir en retour des documents en relation avec le langage de programmation Java, ou avec l'ile Java, ou avec d'autres interprétations auxquelles est liée cette étiquette. Ainsi, la prédiction qui se base sur de tels termes peut induire à des résultats non pertinents qui ne correspondent pas aux attentes des utilisateurs. Nous citons la prédiction fondée sur l'usage des règles d'association qui sont construites à base des historiques d'annotations des utilisateurs, chemin sur lequel nous nous sommes orientés dans ce travail de recherche. Cette prédiction peut induire à i) la sélection de règles non pertinentes, et ii) à la découverte de mauvaises relations entre les données qui appartiennent aux règles exploitées, d'où l'inférence de mauvais intérêts pour l'utilisateur.

La capacité des utilisateurs à interagir les uns aux autres, à échanger des idées et d'opinions, et à former des réseaux sociaux sur le Web, a amené les systèmes à s'intéresser au niveau d'interaction entre eux et à leurs rôles sociaux dans le système. Ces niveaux d'interaction représentent une information sociale qui peut être exploitée à différentes fins. Notre défi est lié à la technique utilisée pour intégrer cette information dans le but d'améliorer la RI pertinente. Cette thèse aborde donc différents défis soulevés dans le domaine du web social afin de l'exploiter dans le processus de la RI. Elle étudie l'impact et la manière de son intégration dans ce processus pour améliorer la pertinence des résultats.

1.1.8. Utilisabilité de l'interface de recherche

Pour garantir une meilleure expérience de recherche à l'utilisateur, un SRI ne doit pas seulement se soucier du retour pertinent des données, mais aussi de la visibilité de ces données sur l'interface utilisateur. Cette visibilité représente un point crucial des SRIs. Comment les résultats doivent être retournés à l'utilisateur ? Cela amène les systèmes à s'intéresser au type d'interactivité qui devrait être fournie à l'utilisateur (English *et al.* 2002). L'accès aux données par les utilisateurs représente un point essentiel dans l'exploration des données. Il offre une facilité et une simplicité de recherche. Dans ce contexte, le défi qui se présente aux systèmes est lié aux développements de méthodes permettant

d'améliorer cette facilité d'utilisation. A ce propos, on distingue les problèmes relatifs aux modèles de représentation des données qui peuvent influencer sur l'apparence de l'espace de navigation (Hearst 2006).

Par ailleurs, on remarque que la majorité des moteurs de recherche populaires renvoie simplement les résultats sous la forme d'une liste ordonnée de documents sans aucune information qui permet de faire la différence entre les interprétations de la recherche. L'utilisateur peut se perdre dans une grande masse d'information. En effet, même si l'information recherchée est présente dans la liste des résultats, elle n'est pas toujours facilement accessible par l'utilisateur. Ainsi, le problème qui se pose est lié au modèle d'organisation et de présentation des résultats qui facilite l'accès aux informations pertinentes et qui peut répondre aux mieux aux attentes des différents utilisateurs.

I.2. Objectifs

L'objectif principal de notre travail de recherche consiste à développer des approches théoriques et des outils pratiques pour améliorer la pertinence des résultats des SRIs, et faciliter l'exploration de grandes collections de données. Pour cela, nous avons l'idée d'étendre le paradigme de la RI à facettes compte tenu de ses nombreux avantages qui ont pu être apportés aux structures mono dimensionnelles en termes de techniques de représentation et d'exploration des données. L'extension proposée s'appuie sur l'intégration d'un ensemble d'éléments qui vont des connaissances extraites depuis des ressources linguistiques, aux concepts empruntés des réseaux sociaux. Ces éléments visent à la fois à représenter, enrichir et à adapter le contenu des documents web aux besoins informationnels des utilisateurs formulés par des requêtes de recherche, et vice versa. Afin de permettre le bon déroulement de cette étude, le développement de cette solution se fait en plusieurs étapes principales que nous considérons, allons du plus général au plus spécifique, dont chacune s'assigne un certain nombre d'objectifs. Ces objectifs sont illustrés à travers la figure 1.1 qui représente une vue globale sur les différents modèles proposés dans cette étude où chacun est lié à un ensemble de limitation qui ont été soulevées dans la section précédente, et à un ensemble d'objectifs que cette thèse vise à atteindre pour combler ces limites.

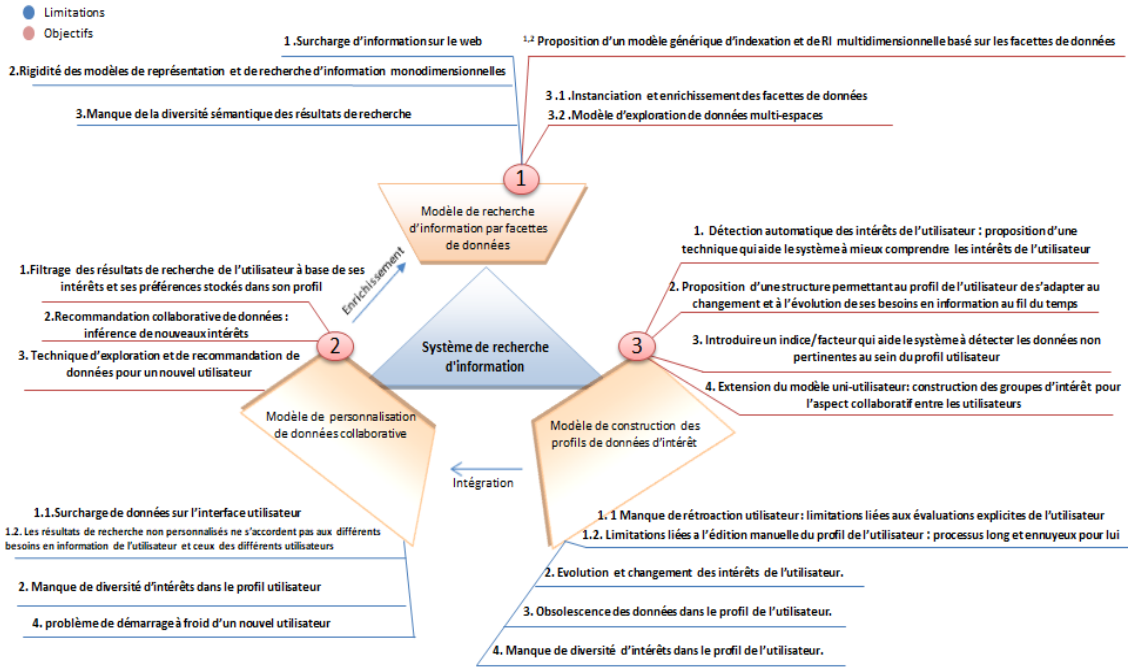


Figure 1. 1. Vue globale sur les limitations soulevées et les objectifs définis dans la thèse

I.2.1. Développement d'un modèle théorique étendu pour la représentation des facettes de données

Nous visons en premier temps à définir un formalisme théorique et générique d'indexation et de RI, qui peut être réutilisable dans plusieurs modèles et différents domaines de la RI. Ce formalisme se base sur la définition de facettes de description qui visent à enrichir l'univers de représentation et de navigation du contenu et à apporter une meilleure flexibilité de recherche.

I.2.2. Modèle uni-utilisateur de profils de données d'intérêt

En réponse à une requête utilisateur, un SRI peut retourner un grand nombre de données. Cependant, l'utilisateur n'est pas nécessairement intéressé par toutes ces données, il doit filtrer manuellement celles à qui il porte plus d'intérêt pour écarter l'information non pertinente. Afin de filtrer automatiquement les résultats selon les intérêts et les préférences de l'utilisateur, un modèle de profil de données d'intérêt est proposé. Ce modèle a pour objectif de réduire l'espace de recherche d'une part, et de rendre exploitables d'une autre part les données d'activités des utilisateurs. Ces données peuvent être utiles pour augmenter la pertinence des résultats et accélérer le processus de RI. Cette proposition vise à combler les lacunes qui

sont liées directement au processus de construction du profil utilisateur, et indirectement lors de son exploitation dans le processus de personnalisation du contenu. Plus concrètement, nous visons à:

- Détecter automatiquement les intérêts de chaque utilisateur par l'introduction d'une technique qui aide le système à comprendre ses intérêts.
- Proposer une structure qui permet au profil de s'adapter au changement et à l'évolution à court et à long terme des intérêts de chaque utilisateur.
- Introduire un indice qui aide à détecter les données non pertinentes dans le profil de l'utilisateur, pour faire face à i) l'obsolescence de ses intérêts au fil du temps qui peut affecter le processus de personnalisation, et face ii) à la surcharge de données dans le contenu du profil, qui peut rendre difficile son exploitation au sein du processus de recherche.

I.2.3. Extension du modèle de profil de données d'intérêt : passage d'un modèle uni-utilisateur à un modèle collaboratif

Grâce à l'émergence des sites de partage de données et des réseaux sociaux tels que Facebook, LinkedIn, Flickr, YouTube, Twitter, delicious, MyFeedz et Google Plus, le web d'aujourd'hui est devenu extrêmement interactif et social. Afin de tenir compte de cette réalité, nous développons un modèle collaboratif qui étend le modèle uni-utilisateur proposé. Le concept central derrière ce modèle est la collaboration entre des utilisateurs qui partagent les mêmes centres intérêts, c'est à dire, aider l'utilisateur courant à bénéficier des expériences des autres utilisateurs similaires. Cet aspect collaboratif servira donc à la formation de groupes d'intérêts derrière les activités similaires, reflétant des communautés virtuelles d'utilisateurs qui aident à personnaliser le contenu en considérant à la fois les caractéristiques individuelles des utilisateurs et celles des groupes d'intérêts.

I.2.4. Développement d'une approche de personnalisation de données

L'approche de personnalisation permet de mieux répondre aux besoins en information de l'utilisateur, en intégrant d'autres paramètres d'accès à l'information en dehors de ses requêtes de recherche. Ces paramètres peuvent influencer positivement sur ce processus de recherche et améliore les résultats renvoyés par le système. Nous citons le profil de l'utilisateur qui englobe ses intérêts et ses

préférences sur lequel se base notre approche de personnalisation. Cette direction vise à i) assister les utilisateurs dans leurs recherches spécifiques d'information à travers un système de filtrage de données, et ii) à suggérer d'autres intérêts que ces utilisateurs n'ont pas encore considérés, à travers un système de prédiction/recommandation de données. Dans notre étude, le filtrage de données consiste à offrir à un utilisateur une vue personnalisée du contenu en se basant sur son profil et sur ceux des autres utilisateurs similaires de ses groupes d'intérêts. Tandis que la recommandation de données permet de suggérer du contenu qui soit lié à d'autres intérêts que l'utilisateur n'a pas encore considérés en se basant sur ce qui a déjà été considéré par lui et/ou par les autres utilisateurs similaires. Cette recommandation aide à enrichir à la fois les connaissances de cet utilisateur et son profil, et fait face au problème manque de diversité d'intérêts dans ce profil.

Tel qu'il a été déjà soulevé plus haut dans la problématique, certaines situations s'avèrent être difficiles pour le système et le rendent incapable de produire aux utilisateurs du contenu personnalisé, en particulier lorsque ce système n'a pas assez d'informations sur les intérêts et les préférences de ces utilisateurs. Ce problème est plus intensif lorsqu'un utilisateur utilise le système pour les premières fois. Notre challenge est de proposer, dans de telles situations, une solution qui permet d'aider cet utilisateur avec du contenu qui peut l'intéresse sans solliciter son aide pour une édition manuelle de ses intérêts. Une méthode interactive de détection d'intérêts et de préférences d'un nouvel utilisateur est proposée. Cette solution vise d'un côté à comprendre les intérêts de cet utilisateur qui permettent au système de lui fournir du contenu pertinent et de construire son profil, et permet d'un autre coté d'enrichir et faire évoluer ses connaissances et ses intérêts.

I.3. Méthodologie et démarche scientifique

Pour atteindre nos objectifs, nous avons développé graduellement une solution d'indexation et de recherche d'information basée sur le concept de multi-facettes en intégrant à chaque étape de nouveaux aspects tels que l'analyse et l'interprétation du contenu informationnel, la représentation et l'enrichissement des facettes, l'intégration des concepts d'extraction et de représentation des caractéristiques utilisateurs qui sont relatives à leurs activités de recherche et à leurs rôles sociaux dans le

système, et l'intégration des approches de fouille et de personnalisation de données basées sur l'implication collective de ces utilisateurs. Une vue globale de ces aspects est résumée dans la figure 1.2.

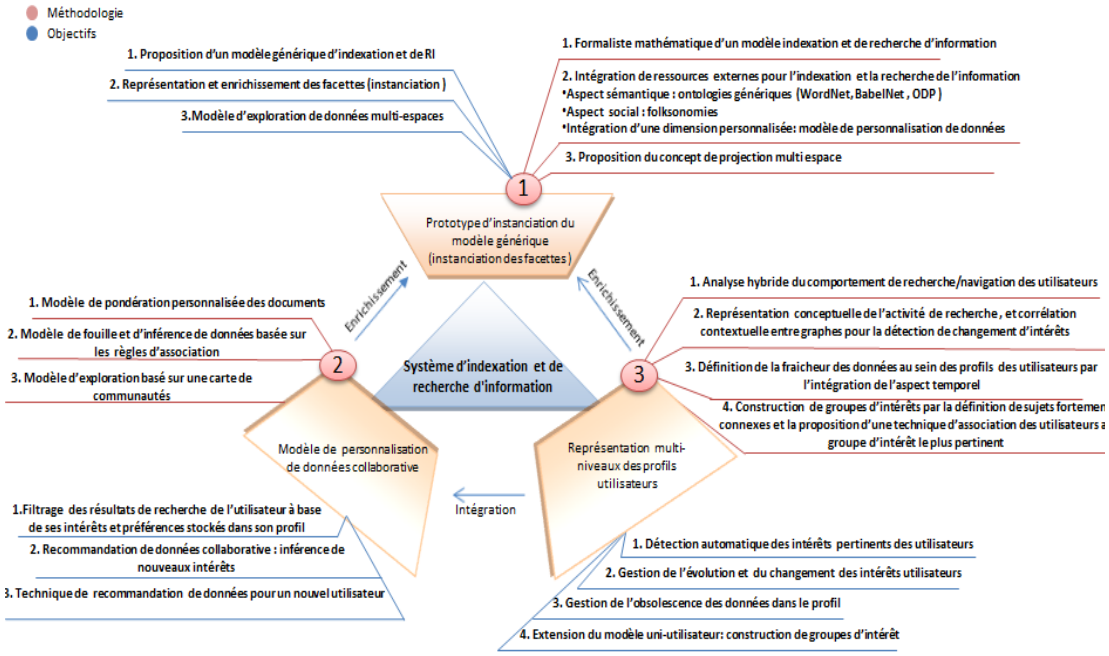


Figure 1. 2. Vue globale sur les méthodologies adoptées pour atteindre les objectifs de la thèse

I.3.1. Développement d'un modèle étendu pour la représentation des facettes

Nous avons choisi d'utiliser un langage mathématique pour la définition de notre modèle générique, il se base pour la représentation des facettes sur le nouveau concept de projection multi-espaces. L'adoption d'un modèle mathématique est justifié par le fait qu'un tel langage se caractérise par un mode d'expression formalisé, précis, et rigoureux, et son pouvoir de faire abstraction de la sémantique, ce qui rend nos théories réutilisables dans plusieurs modèles et domaines de la RI.

Le formalisme proposé est ensuite instancié et mis en œuvre à travers un prototype qui exploite des données réelles de différentes sources de connaissance. Ce prototype a pour objectif d'évaluer l'apport du modèle théorique proposé. Pour ce faire, nous intégrons pour la représentation de nos facettes : i) des connaissances linguistiques via des ontologies lexicales et des dictionnaires linguistiques tels que BabelNet[1] et Wikipédia[2], ii) des connaissances du domaine à travers des ontologies génériques telles que le répertoire des pages web ODP, et iii) des connaissances sociales telles que les folksonomies et des

[1] BabelNet : http://babelnet.org/
[2] Wikipedia : http://en.wikipedia.org/wiki/Wikipedia_database

modèles utilisateurs qui représentent et constituent les caractéristiques spécifiques des utilisateurs, en particulier leurs intérêts et préférences, et leurs rôles sociaux. Ces bases de connaissance sont utilisées pour représenter, organiser, et enrichir l'univers d'indexation des documents, ainsi que l'univers d'interprétation des requêtes utilisateurs et de leurs intérêts. Chacune est exploitée pour préparer un espace de représentation. Nous comptons de cette façon introduire le concept de facettes multiples basé sur les vues de données que nous appelons par « v-facettes » où chaque vue est incarnée par un espace de données qui représente une structure de description différente. Cet enrichissement permet de réduire l'écart existant entre l'univers de représentation des documents et celui des requêtes utilisateur. En se basant sur ce nouveau modèle de RI, une approche est proposée pour permettre aux utilisateurs d'explorer les résultats des recherches suivant les facettes de données construites et des valeurs de facettes.

Afin de bien mener cette étape d'enrichissement, l'analyse du contenu des documents est la première démarche prise en compte. Cette étape est nécessaire autant pour les documents que pour les requêtes des utilisateurs. Elle est caractérisée par la définition d'un ensemble de techniques nécessaires pour i) la transformation du contenu web en une forme compréhensible et accessible par le système, ainsi que pour ii) la bonne interprétation du besoin utilisateur et sa transformation en une forme plus exploitable par le processus de recherche. Ensuite, le SRI est mis en œuvre par la mise en correspondance entre les deux univers de représentation préparés. Cette correspondance est basée sur le concept de couverture de données. Elle est complétée par une technique de navigation et d'exploration de données.

I.3.2. Introduction d'un modèle utilisateur de profils de données d'intérêt et de groupes d'intérêt

La modélisation du profil de l'utilisateur consiste à l'énumération des données qui représentent ses caractéristiques spécifiques pouvant impacter positivement la RI, puis la définition d'un modèle approprié pour représenter ces données. L'objectif est la mise en place d'une structure exploitable au sein d'un processus de personnalisation et efficace pour retourner du contenu pertinent. Dans le domaine de la RI personnalisée, le profil de l'utilisateur décrit le plus souvent ses centres d'intérêt pouvant être représentés de différente manières. Dans notre étude, la construction du profil utilisateur se base sur l'exploitation des données qui sont issues de ses interactions avec le système, connues sous le nom des données

d'activités. Dans notre modèle, ces données peuvent être des requêtes de recherche, des étiquettes d'annotation, des documents jugés implicitement ou explicitement pertinents par l'utilisateur cible, ou des facettes d'intérêt. Elles représentent les données d'intérêt spécifiques de l'utilisateur, que nous appelons aussi par les données d'intérêt granulaires, et font l'objet de construction de plusieurs niveaux de représentation abstraits qui permettent de modéliser ces données spécifiques sous différents aspects de représentation génériques. Le premier niveau abstrait comporte les activités de recherche où chacune regroupe les données d'interactions qui sont liées à une même requête de recherche. Le deuxième niveau est l'ensemble des sujets d'intérêt où chacun englobe les activités de recherche qui sont liées à un même besoin en information, ce niveau vise à atténuer le problème de l'ambigüité des termes soulevé avec les étiquettes et les requêtes polysémiques. Le dernier niveau de représentation englobe les groupes de sujets fortement connexes qui visent à construire des communautés d'intérêts à base de sujets similaires.

Afin de gérer au mieux l'exploitation des données d'intérêt de l'utilisateur depuis son profil, trois critères de pertinence sont utilisés, à savoir, le contexte, la fraicheur et la fréquence des données. La fraicheur de données est utilisée pour faire face à l'obsolescence des intérêts de l'utilisateur au fil du temps. Cela est rendu possible par l'annotation temporelle des données d'activités de l'utilisateur. Cette annotation permet de déduire leur fraicheur à chaque période de temps et vise à éliminer les données non pertinentes pouvant affecter le processus de personnalisation. Le contexte aide à identifier les données d'intérêt qui couvrent un besoin informationnel donné. Enfin, la fréquence d'utilisation aide à définir les données d'intérêt récurrentes. Une technique de combinaison de ces trois facteurs influents est proposée.

Ensuite, le modèle uni-utilisateur est étendu par la proposition d'une technique qui permet d'extraire les ressemblances comportementales entre les utilisateurs. Ceci aide à la préparation de la méthode de filtrage collaboratif (FC). Pour ce faire, la similarité entre les profils de ces utilisateurs est évaluée. Elle prend en compte deux niveaux de comparaison :

- Niveau de comparaison générique qui compare les utilisateurs à travers leurs sujets d'intérêt.
- Niveau de comparaison spécifique qui compare ces utilisateurs à travers leurs données d'activités spécifiques, et prend en compte également la fraicheur de ces données.

L'hypothèse derrière ces deux niveaux de comparaison est que deux utilisateurs peuvent être intéressés par les mêmes sujets d'intérêt (niveau d'intérêt générique) sans avoir consommé les mêmes données

spécifiques (documents, étiquettes, requêtes). Cette hybridation ne pénalise donc pas ces utilisateurs et les considère comme similaires à un niveau générique et peuvent être fructueux les uns aux autres. L'aspect de comparaison spécifique quant à lui permet de renforcer la similarité entre les utilisateurs qui exploitent les mêmes données spécifiques dans la même période de temps. Une méthode de combinaison entre les deux niveaux de comparaison, est proposée.

I.3.3 Exploitation du profil utilisateur

A. Indexation personnalisée des documents

Les requêtes de recherche et les étiquettes d'annotations décrivent les utilisateurs dans leurs profils. Elles décrivent aussi les documents qui ont été découverts par le système suite à leur exploitation par les utilisateurs. En considérant cette hypothèse, nous exploitons ces objets de contenu pour l'enrichissement des documents dans l'index. Cela permet d'apporter à l'univers de représentation de ces documents une dimension de description personnalisée. Cette dimension permet de les décrire selon chaque utilisateur et aide à filtrer les résultats de recherche de ces utilisateurs selon leurs intérêts et préférences. L'idée derrière cette intégration est d'essayer d'impacter le score global de correspondance «document-requête» avec les intérêts de l'utilisateur. Cela est fait en considérant dans la représentation des documents, les termes que les utilisateurs emploient souvent pour exprimer leurs besoins à travers des requêtes de recherche et/ou des étiquettes d'annotation. Il s'agit d'ajuster les scores de pondération des termes dans l'index documentaire quand ils correspondent aux intérêts des utilisateurs. Cette nouvelle pondération permet promouvoir les documents interrogés correspondant à ces intérêts d'enrichissement.

B. Développement d'une approche de personnalisation de données

Nous nous sommes orientés pour le filtrage de données vers une méthode hybride qui repose sur la combinaison linéaire de plusieurs scores d'appariement du contenu à proposer dernière une requête de recherche. Cette hybridation s'appuie sur les expériences menées par (Vogt *et al.* 1996) au sujet de la pertinence de recherche. Ces auteurs ont montré que les résultats de recherche sont significativement améliorés avec une simple combinaison linéaire de scores retournés par différentes approches de RI. Une fonction de personnalisation est proposée, elle permet de déterminer la similarité globale d'un document à retourner à l'utilisateur derrière sa requête de recherche en fonction du contenu de cette requête, de son

profil et des profils de ses utilisateurs voisins qui sont calculés à travers une pertinence multidimensionnelle. Cette pertinence permet d'associer à chaque utilisateur un groupe d'intérêt représenté par un sous-ensemble d'utilisateurs ayant un intérêt pour un ou plusieurs sujets fortement connexes.

La détection des comportements communs entre les utilisateurs peut-être aussi considérée en termes de la ressemblance ensembliste des groupes de données observés chez eux. Ceci aide à prédire les intérêts d'un utilisateur lorsqu'un sous-ensemble de ses intérêts est observé fréquemment chez un groupe d'utilisateurs. Pour ce faire, un mécanisme, qui permet de faire le passage de ces ensembles de données à des prédictions, doit être possible. Afin de mettre en valeur et en évidence cette hypothèse, des règles association sont exploitées pour représenter la dépendance entres les ensembles de données d'intérêt observés fréquemment chez les utilisateurs. Les données de départ sont extraites du contenu des profils utilisateurs et sont utilisées pour extraire les données fréquentes.

Étant donné que le volume de ces règles d'association, et leur exploitation dans le processus de prédiction des intérêts des utilisateurs sont au cœur de notre problématique, des facteurs d'importance sont proposés pour promouvoir les règles les plus pertinentes pour chaque utilisateur. Cela consiste à personnaliser l'usage de ces règles pour chaque utilisateur, et cela en basant sur son affinité envers les données d'intérêt. A ce propos, des métriques d'affinité sont proposées. Elles représentent le degré d'intérêt des utilisateurs envers ces données. Ces degrés d'intérêt sont calculés à travers la fréquence d'utilisation des données, leur popularité, et leur fraicheur dans les profils de ces utilisateurs, etc.

I.4. Mise en œuvre et évaluations

Pour tester le nouveau système proposé, un prototype de moteur de recherche est développé, et une série d'expérimentations est conduite à chaque étape du processus pour évaluer l'apport effectif de chaque modèle proposé, et permettre d'éventuelles améliorations. Pour ce faire, différentes sources de données sont exploitées pour instancier les modèles proposés. Ensuite différentes mesures de pertinence sont utilisées pour l'évaluation de leur efficacité, en allant du modèle le plus général au plus spécifique.

- Proposition des mesures de pertinence de la RI facettée. Ces mesures permettent d'évaluer les systèmes qui sélectionnent les facettes et leurs valeurs pour une interface de recherche facettée.

- Des métriques d'évaluation de la qualité des profils multidimensionnels, sont exploitées, notamment la précision aux premiers X concepts (Daoud 2009) pour l'évaluation du niveau sémantique de ce profil, ainsi que d'autres mesures qui sont liées aux autres niveaux de représentation de ce modèle utilisateur.
- D'autres métriques classiques sont aussi utilisées pour l'évaluation des modèles de personnalisation proposés. Elles sont destinées à l'évaluation de l'efficacité de la RI, à savoir, la précision P@X, la moyenne des précisions MAP (Means Average Precision), et le rappel moyen.

I.5. Originalité

La nouveauté de la contribution de cette thèse provient de l'intégration dans un même système, à travers des objectifs clairs et précis, des éléments provenant de différents domaines, notamment, la recherche d'information, la fouille de données, la personnalisation Web, la représentation des connaissances, et l'analyse des réseaux sociaux. Différents challenges soulevés dans ces différents domaines sont adressés, en particulier lorsque ces différents aspects sont exploités dans la RI en vue d'améliorer sa pertinence. Du point de vue pratique, les résultats de cette recherche pourraient être exploités pour non seulement améliorer les SRI existants, mais aussi pour créer de nouveaux services dans l'objectif est de mieux assister les utilisateurs du Web dans leur quête d'informations pertinentes.

I.6. Plan de thèse

Ce chapitre a présenté le contexte de notre travail de recherche ainsi que les principaux défis qui restent encore soulevés dans la littérature que nous proposons de couvrir dans cette thèse. Le reste de cette thèse est divisé en six chapitres : le chapitre 2 présente les notions et les concepts de base de la RI et l'évolution des techniques qui sont proposées dans ce domaine, en commençant par les approches classiques aux approches adaptatives qui intègrent différentes connaissances externes, telles que les connaissances du domaine pour le cas des SRIs sémantiques, des paramètres contextuels pour la recherche contextuelle, et l'aspect social pour les SRIs personnalisés et sociaux. Ce chapitre inclut aussi les avantages et les inconvénients qui sont associés aux différentes approches discutées. Les autres chapitres sont consacrés à nos contributions. Dans le chapitre 3, nous proposons un modèle générique de

RI par facettes. Ensuite dans le chapitre 4, un modèle générique de profil utilisateur est proposé, il prend en compte différents aspects de personnalisation. L'intégration de ce modèle utilisateur dans le processus de recherche fait l'objet de deux systèmes de personnalisation, à savoir, un système de filtrage de données et un système de prédiction d'intérêts. Ces propositions sont discutées dans les deux chapitres 4 et 5. Ensuite, un modèle qui adresse le problème de démarrage à froid d'un nouvel utilisateur est proposé dans le chapitre 6. Enfin, le système proposé dans cette étude est évalué dans le chapitre 7. Cette évaluation commence par l'instanciation des modèles proposés en exploitant des données réelles permettant d'alimenter le système. La validation consiste ensuite à tester les résultats obtenus de chacun de ces différents modèles selon des métriques différentes et adéquates citées dans la section 1.4. Nous terminons par une conclusion générale et des perspectives de recherche dans le chapitre 8.

Chapitre 2: Recherche d'information classique et émergence des approches avancées

II.1. Partie 1 : Recherche d'information classique

II.1.1. Introduction

Chercher une information sur le web devient aujourd'hui une activité courante et nécessaire pour presque tout le monde. Ce domaine de recherche s'intéresse à extraire et organiser l'information pour pouvoir facilement la chercher plus tard. L'objectif principal derrière ce processus est de sélectionner les informations qui répondent aux mieux aux besoins des utilisateurs en intégrant différents modèles et techniques d'accès à l'information (Salton et McGill 1986; Wang *et al.* 2009). Cette recherche d'information (RI) peut-être active, appelée aussi la recherche classique, lorsque l'accès à l'information s'effectue uniquement à travers la requête utilisateur. Elle peut-être passive si d'autres ressources externes en dehors de cette requête sont exploitées, comme les documents jugés, les termes pertinents, ou encore des connaissances du domaine. Ces ressources sont utilisées pour la représentation et l'enrichissement du contenu documentaire, et/ou l'expansion du contenu de cette requête qui est souvent mal exprimé par les utilisateurs et ne reflète pas toujours leur besoin réel. Dans ce cas la recherche est considérée comme adaptative. L'accès à l'information peut être aussi effectué à travers un modèle utilisateur qui décrit ses caractéristiques spécifiques (ex. données personnelles, données géographiques, préférences, centres d'intérêt, interactions de recherche, ou d'autres caractéristiques de son environnement). Cette direction de recherche est nommée par « la recherche personnalisée », elle permet d'adapter la recherche à un contexte bien précis et particulier de l'utilisateur.

Avec la croissance d'internet, différents types de réseaux sociaux (RS) se sont formés, ils représentent aujourd'hui le moyen le plus utilisé pour la diffusion et le partage de ressources et de connaissances sur le Web. Les utilisateurs sont alors confrontés à une grande quantité d'information qui contribue au développement de leurs connaissances, et ils se retrouvent conséquemment avec de nouveaux besoins en information. Cette thématique a permis l'adaptation de la recherche dans un nouveau contexte appelé le contexte social, en considérant les informations spécifiques aux réseaux sociaux (RS),

telles que les annotations sociales et/ou relations sociales, ou/et les rôles des utilisateurs dans le système. De nombreux travaux ont abordé l'exploitation de ces informations dans les systèmes de personnalisation de données notamment les systèmes de recommandation et les systèmes de filtrage collaboratif.

Ce chapitre a pour objectif de porter la lumière sur le domaine de la RI. Nous commençons par donner quelques concepts et définitions de base qui sont liés au SRI et à ses principales fonctions. Nous énonçons ensuite différentes modalités de recherche et d'exploration qui peuvent être offertes aux utilisateurs à travers une interface de recherche. Nous passerons ensuite en revue les principaux modèles existants en accès à l'information en allons des premières générations classiques jusqu'aux approches avancées (adaptatives et personnalisées). Cette nouvelle génération vise à atténuer les problèmes causés par une recherche classique en vue de l'amélioration de la qualité des résultats. En dernier, nous finirons par une analyse des avantages et des inconvénients inhérents à ces approches, permettant de positionner notre travail dans la littérature.

II.1.2. Système de recherche d'information

II.1.2.1. Concepts de base et définitions

Un SRI est un ensemble de techniques qui assurent les fonctions nécessaires pour la RI. Il a pour rôle de sélectionner les documents qui peuvent répondre au besoin en information de l'utilisateur formulé par une requête de recherche. Cette requête est souvent sous forme d'un langage naturel, c'est-à-dire un ensemble de mots clés, la requête est considérée dans ce cas comme basique, ou peut être sous une composition de requêtes basiques avec un ensemble d'opérateurs logiques (ET, OU, SAUF, etc.) appelée requête logique (Jansen *et al.* 2000), ce type de requêtes est souvent utilisé dans les modèles classiques, à noter les modèles booléens (Salton *et al.* 1983) (Salton 1989). La requête peut aussi être formulée sous un langage structuré de requêtes tel que SPARQL pour les données RDF (Corby *et al.* 2004) (Corby *et al.* 2006) ou le langage graphique pour les documents XML (Ykhlef et Alqahtani 2011). L'information recherchée quant à elle peut être un texte, un morceau de texte, une page web, une image, une vidéo, ou toute autre ressource d'information répondant à un besoin informationnel.

L'objectif principal de tout SRI est de garantir aux utilisateurs une meilleure pertinence de résultats. Tel qu'il a été soulevé dans le chapitre précédent, cette pertinence est au cœur des défis des SRIs, notamment pour faire face à la diversité et la quantité d'information disponible sur le web. Elle mesure la correspondance entre les documents et une requête pour retourner uniquement ceux qui répondent au mieux au besoin de l'utilisateur. Ce concept ne se limite pas à cette définition, plusieurs travaux de recherche s'accordent sur la difficulté de sa définition et ils se sont intéressés à lui trouver une définition formelle (DAFT et HUBER 1975) (Mizzaro 1997) (Borlund et Ingwersen 1998). D'une manière générale, deux types de pertinence ont été distingués dans la littérature (Karbasi 2007) (Aouicha 2009) : la pertinence système et la pertinence utilisateur.

- La pertinence système est définie à travers les modèles de RI. Elle est souvent traduite comme étant un score évaluant la correspondance du contenu d'un document vis-à-vis de la requête utilisateur (Cleverdon 1970). Ce score est généralement évalué en fonction des poids des mots de la requête dans le document interrogé. Ces poids représentent l'importance des mots pour le contenu d'un document.
- La pertinence utilisateur quant à elle, est liée à l'évaluation effectuée par l'utilisateur en ce qui concerne l'information renvoyée par le système (Harter 1992) (Saracevic 1996) (Borlund et Ingwersen 1998). Les techniques de retour de pertinence trouvent leur origine dans les années 1970 avec notamment les travaux de Rocchio sur le retour explicite de pertinence (feedback utilisateur) (Rocchio 1971). Les évaluations de l'utilisateur peuvent être aussi implicites ou explicite, et elles peuvent être exploitées dans la construction de son profil qui sera ensuite intégré dans le processus de la RI personnalisée, ces notions sont abordées avec plus de détails dans la section II.2.3 de ce chapitre.

L'évaluation de cette pertinence a donné aussi lieu à d'autres études qui ont défini un ensemble de critères permettant de déterminer la pertinence d'un contenu retourné à un utilisateur donné selon un contexte de recherche particulier. La pertinence d'un document varie donc d'un utilisateur à l'autre, et chez le même utilisateur d'un contexte à l'autre. En considérant cette variation, des facteurs pouvant affecter les jugements de pertinence des utilisateurs ont été identifiés, ceci a fait l'objet de plusieurs travaux (Cuadra et Katter 1967), (Cuadra et Katter 1967) (Barry 1994). Dans (Barry 1994), ces facteurs

sont classés en six catégories : (a) le contenu informationnel des documents (b) les sources des documents (c) l'aspect physique des documents (d) les préférences de l'utilisateur (e) le niveau d'expertise et de connaissance de l'utilisateur (f) les informations en relation avec son environnement. En fonction de ces facteurs, plusieurs pertinences sont possibles entre un document et une requête (Tamine et Calabretto 2008) (Daoud 2009), parmi lesquelles on note :

- **La pertinence algorithmique** : il s'agit de la pertinence système, elle se base sur le calcul de correspondance du contenu documentaire par rapport à celui de la requête en considérant les caractéristiques des documents d'une part et celles des requêtes d'une autre part. Cette pertinence est donc complètement indépendante de l'utilisateur.
- **La pertinence thématique** : elle est définie par le niveau de couverture du contenu de document pour le thème évoqué par une requête de recherche. Cette pertinence est exploitée dans les campagnes d'évaluation telle que TREC (Harter et Hert 1997).
- **La pertinence contextuelle** : appelée aussi la pertinence situationnelle, elle est dynamique et dépend du contexte de recherche de l'utilisateur et de sa perception.
- **La pertinence cognitive** : elle est liée aux connaissances et à la perception de l'utilisateur envers un thème de sa requête. Elle est dite cognitive car elle permet d'améliorer la connaissance de l'utilisateur via le contenu renvoyé au cours de sa recherche.

Un SRI performant est donc celui qui supporte un modèle de RI qui rapproche le plus possible la valeur de sa pertinence aux jugements de pertinence donnés par les utilisateurs.

Cette section a permis de présenter les notions de base les plus fréquemment utilisées dans le domaine de la RI. La prochaine section présente le fonctionnement général d'un SRI que nous jugeons nécessaire à comprendre pour la compréhension du reste du manuscrit.

II.1.2.2. Fonctionnement du système de recherche d'information

Pour répondre à une requête utilisateur, un SRI met en œuvre un certain nombre de processus pour réaliser la mise en correspondance entre le contenu des documents web d'une part, et celui de la requête utilisateur d'une autre part. Il est défini par ses modèles de représentation des documents et des requêtes utilisateur, et sa fonction de recherche pour la mise en correspondance entre les deux univers de représentations. Ce processus est composé de deux fonctions principales :

1. **Modèle de représentation**: le processus de prétraitement du contenu consiste en une étape très importante pour un SRI, car de sa qualité dépendra la pertinence des résultats, il est dédié pour la représentation du contenu des documents et du besoin informationnel de l'utilisateur formulé sous une requête de recherche.

   a) **Le prétraitement des documents**: il consiste à extraire à partir des documents une représentation qui couvre au mieux leur contenu. Cette opération est connue aussi sous le nom de l'interprétation ou l'analyse du contenu. Elle consiste à l'extraction d'un ensemble de descripteurs les plus représentatifs du contenu, ces descripteurs sont appelés aussi par les entrées de l'index ou les termes d'indexation, utilisés pour l'indexation de ces documents. Cet index documentaire est similaire au principe des index utilisés dans les livres pour faciliter l'accès à leur contenu. Un descripteur dans l'index documentaire peut-être de différentes catégories : mot simple ou composé, lemme (origine lexicale du mot), N-gramme (séquence de N caractères consécutifs), concept (on parle alors d'indexation conceptuelle), etc. L'indexation est considérée i) manuelle quand les documents sont analysés par un spécialiste ou un expert du domaine, ii) automatique lorsque l'analyse se fait à l'aide d'un processus automatique, ou iii) semi-automatique quand les deux méthodes sont combinées. En général l'ensemble de traitements effectués sur le contenu des documents durant un processus d'indexation automatique comprend: l'extraction des mots simples, l'élimination des mots vides (les mots les moins pertinents tels que les déterminants), la lemmatisation qui consiste à normaliser les descripteurs par leur lemme, et la pondération des descripteurs qui consiste à évaluer leur importance dans le document. Celle-ci dépend de plusieurs propriétés (ex. la fréquence du terme dans le document (TF), sa fréquence dans la collection (DF), la taille du document, etc.), qui sont identifiées selon le modèle d'indexation, les plus connus sont: le modèle IF-IDF (Sparck Jones 1972) (Dillon 1983), la loi de Zipf (Zipf 2016), et le modèle probabiliste BM25 (Robertson *et al.* 1995) et ses extensions BM25F (Robertson *et al.* 2004) et BM25t (Géry *et al.* 2010), BM25F_S (Bouhini *et al.* 2013a).

En fonction des ressources exploitées lors de l'indexation des documents, cette dernière peut être catégorisée en deux types : indexation libre, et indexation contrôlée. Dans la première indexation, les descripteurs sont choisis librement sans dépendre à un vocabulaire de termes prédéfinis. En instance, nous citons l'indexation conceptuelle à base de cooccurrence des termes (Koll 1979) (Lipczak 2008), ou l'indexation sociale dans les folksonomies qui se base sur des termes libres et spontanés employés par des utilisateurs non spécialistes (Auray 2007); tandis que dans la deuxième catégorie, les descripteurs sont prédéfinis dans une ressource de référence tels qu'un thésaurus, une terminologie, ou une ontologie.

b) **L'interprétation des requêtes des utilisateurs** : cette opération a pour rôle de représenter le besoin en information des utilisateurs. Comme c'est le cas avec les documents, il s'agit d'extraire les descripteurs les plus représentatifs du contenu de la requête en se basant sur une analyse qui peut couvrir une ou plusieurs dimensions (syntaxique, lexicale, sémantique, etc.). Afin de mieux exprimer le besoin en information de l'utilisateur, le contenu de cette requête peut être reformulé par la génération d'une nouvelle requête, censée être plus précise et appropriée, à partir de son contenu initial, ou elle peut être étendue avec des termes supplémentaires pouvant être de différentes natures (sémantique, sociale, contextuelle, etc.).

2. **Modèle de recherche ou correspondance requête-document** : selon les représentations de la requête et des documents, le SRI effectue un appariement entre ces deux univers de représentation, en vue d'évaluer la pertinence des documents vis-à-vis de la requête. Le système décidera si un document est pertinent, et le sélectionnera pour le présenter à l'utilisateur, c'est ce que a été défini par la pertinence du système. Cet appariement peut-être exact tel est le cas avec les modèles booléens dans lequel les documents résultants ont tous la même pertinence et ne sont donc pas triés (cf. section II.1.4.1). Il peut être aussi approximatif dans lequel les documents résultants peuvent être ordonnés selon le degré de pertinence vis-à-vis la requête (cf. section II.1.4.2). Cette valeur de pertinence est calculée à partir d'une probabilité ou une similarité appelée en anglais « Retrieval Status Value » et est notée RSV (q, d), où « q » est une requête et « d » un document. Cette mesure tient compte des

poids des termes dans le modèle de représentation défini pour les documents et les requêtes (Imafouo et Tannier 2005).

Comme nous l'avons évoqué plus haut, dans certains modèles l'utilisateur peut indiquer au système si le document répond ou non à ses attentes. Les évaluations de l'utilisateur appelées aussi le retour de l'utilisateur aident le système à reconstruire automatiquement une nouvelle requête, afin de présenter une nouvelle liste de résultat. À ce propos, on trouve les techniques de réinjection de pertinence (Rocchio 1971) (Salton et Buckley 1997) (Lavrenko et Croft 2001), qui consistent à reformuler ou à enrichir les requêtes utilisateurs avec d'autres informations additionnelles en se basant sur le retour des utilisateurs, ce qui rend le processus itératif et aide à améliorer la pertinence des résultats (on parle alors de la pertinence cognitive abordée plus haut). La recherche est considérée comme adaptative. Dans le but d'augmenter la performance de la recherche, le système devra réagir assez rapidement à son utilisateur. Nous illustrons, à travers la figure 2.1, l'architecture générale d'un SRI selon les chercheurs (Salton et McGill 1986).

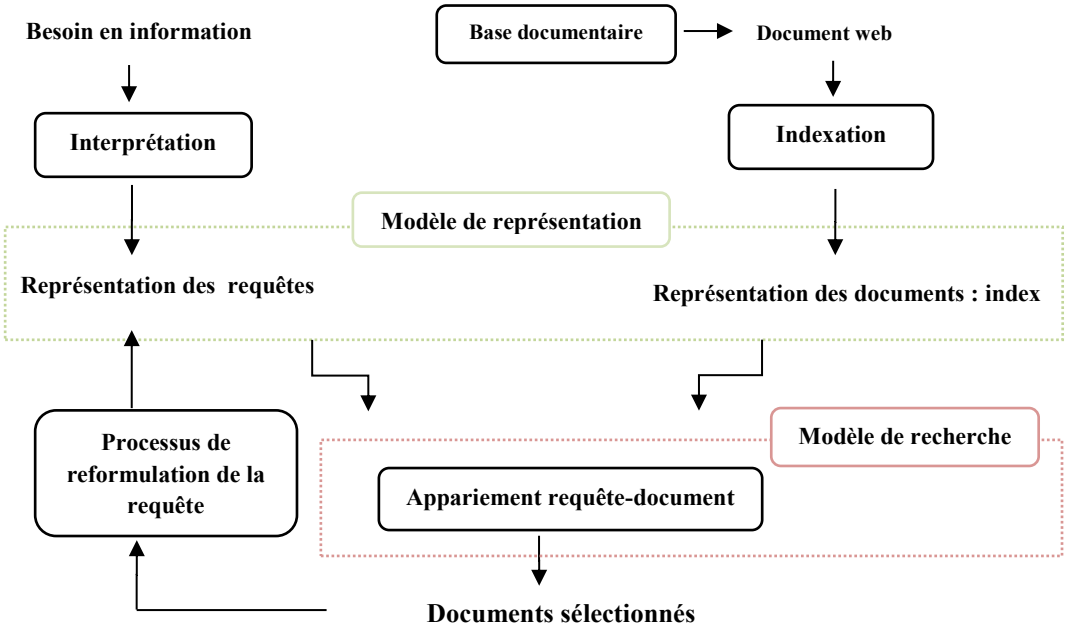


Figure 2. 1. Architecture d'un SRI selon (Salton et McGill 1986)

II.1.3. Stratégies de recherche

II.1.3.1. Recherche par mots clés

Ce mode de recherche permet à l'utilisateur d'exprimer son besoin en information sous la forme d'un ensemble de mots clés à travers une boite de recherche, et les résultats de recherche sont exprimés uniquement dans une liste de résultats. Plusieurs modèles de RI sont proposés dans la littérature dans lesquels les utilisateurs peuvent poser des requêtes constituées de mots-clés (cf. section II1.4.4.). Google, Yahoo et Bing et d'autres moteurs de recherche grand public adoptent cette approche. Pour établir cette recherche, le système sélectionne les documents qui correspondent à un ou plusieurs mots de la requête utilisateur dans l'index documentaire. Cette sélection peut être simple lorsque l'appariement requête-document se base sur le contenu exact de la requête et des documents interrogés. Il peut être fondé sur un enrichissement lorsque le contenu est étendu à travers des ressources externes. Cette extension aide à améliorer la performance des systèmes traditionnels en exploitant des connaissances du domaine (le cas de la recherche sémantique), ou des informations contextuelles relatives à l'utilisateur (la recherche contextuelle), ou des ressources sociales telles que les annotations employées par les utilisateurs sur les documents, les relations sociales entre utilisateurs (recherche sociale), ou à travers des informations reflétant les intérêts et les préférences des utilisateurs stockées dans un profil (le cas de la recherche personnalisée). Ces informations permettent de rapprocher les informations entre elles lors du processus de recherche en adaptant la recherche selon un paramètre de performance sémantique, contextuel, ou social. Elles peuvent être intégrées à différent niveau du processus de la recherche, défini par le modèle en vue d'améliorer les résultats. Ces directions de recherche sont abordées avec plus de détail dans la section II.2.

II.1.3.2. Recherche par navigation

Parmi les modalités de recherche existantes, on trouve les systèmes de recherche par navigation (Godin *et al.* 1993) (Godin *et al.* 1993) (Hascoët et Beaudouin-Lafon 2001). Il s'agit d'une recherche exploratoire qui permet à un utilisateur qui n'a souvent pas de connaissances préalables sur les éléments informationnels du système, de les découvrir selon ses besoins, et cela en lui proposant des critères d'exploration. À chaque étape de la navigation, l'utilisateur n'a qu'à faire un ou plusieurs choix parmi un

ensemble d'alternatives qui lui est proposé par le système. La combinaison entre les deux modes de recherche (la recherche par mots clés et exploratoire) donne lieu à ce qui est couramment appelé par l'approche hybride. Elle est conçue pour les utilisateurs novices qui n'ont pas assez d'expériences sur le domaine de leur recherche, ou qui ne savent pas exactement quoi chercher, le tout sans pénaliser les utilisateurs qui savent choisir les mots clés appropriés pour exprimer leurs besoins informationnels et savent où ils doivent se rendre dans l'espace d'exploration.

Il existe plusieurs stratégies de navigation, nous citons les systèmes de navigation par taxonomies qui sont souvent utilisés. Ils organisent les données sous un arbre hiérarchique, et s'appuient sur un mécanisme de recherche basé sur les relations (générales et spécifiques) liant les données dans l'arbre taxonomique (Amar 2009). Pour représenter ces données, différentes métadonnées peuvent être exploitées, telles que les ontologies exprimées sous un langage formalisé (RDF/XML, RDFS, OWL, SKOS, etc.), des thesaurus, les cartes de domaines (dites en anglais topic Maps) (Pepper et Garshol 2002), etc. Dans ce cas on parle de la « navigation par métadonnées ». Elle permet d'étendre les fonctionnalités des affichages de liste en permettant aux utilisateurs de rechercher dans une hiérarchie de données de manière à obtenir un sous-ensemble basé sur un filtre de navigation. Tel que cela a été soulevé dans le chapitre précédent, les modèles de représentation qui se basent sur de telles ressources souffrent de centaines limitations qui peuvent affecter l'accès aux informations, cette faiblesse peut rendre l'exploration de données lente et contraignante.

On distingue également les systèmes folksonomiques qui permettent l'exploration du contenu des ressources à travers des nuages d'étiquettes (Roxin et Bernard 2007). Ces nuages offrent une vue globale sur les étiquettes qui ont été appliquées à des ressources par différents utilisateurs et offrent différentes représentations visuelles de leur contenu. Différents paramètres peuvent être définis et appliqués sur les étiquettes pour un affichage personnalisé du contenu. L'utilisateur peut par exemple afficher toutes les étiquettes disponibles ou les afficher par fréquence d'utilisation et nombre d'occurrences qui représentent leur popularité (Zubiaga *et al.* 2009). Les générateurs de nuages se basent aussi sur les relations qui existent entre les termes (synonymie, variations lexicales, etc.) (Cheng *et al.* 2014).

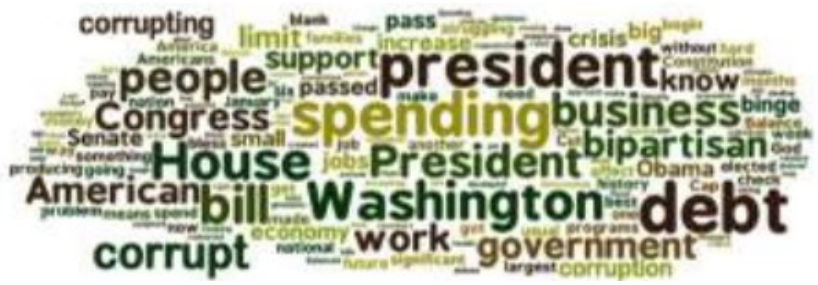


Figure 2. 2. Exemple d'un nuage d'étiquettes

II.1.3.3. Recherche facettée

Pour offrir plus de flexibilité aux utilisateurs, le concept de facettes a vu son apparition à travers un mécanisme de navigation multidimensionnelle (Group et Vickery 1963). Il s'agit d'un ensemble de métadonnées qui définissent des catégories dans le contenu informationnel ou une partie de celui-ci, et utilisées pour regrouper les éléments de ce contenu. Si nous considérons un SRI où les documents sont écrits soit en français, en anglais ou en arabe, lors des recherches, les documents retournés peuvent être regroupés suivant la langue. Dans ce cas, la langue est une facette de description, et les valeurs « français », « anglais » et « arabe » sont des valeurs de cette facette. Contrairement au principe de la classification hiérarchique, qui ne permet d'attribuer qu'une seule dimension à la fois pour décrire l'information, et qui répond généralement à une taxonomie prédéfinie, la classification par facettes permet d'identifier plusieurs critères définis selon le besoin pour décrire l'information (Atanassova et Bertin 2014) (Gödert 2014). Cela permet de rendre l'accès à l'information plus facile et plus rapide, et offre une flexibilité en combinant par exemple plusieurs facettes pour filtrer rapidement un sous-ensemble de contenu qui intéresse l'utilisateur. En outre, chaque utilisateur peut choisir l'ordre des filtres selon ses besoins. À titre d'exemple, nous citons le cas d'un système de classification de restaurants par facettes, où un utilisateur u_1 peut lancer sa recherche suivant le prix, puis l'affiner par localisation et par type de cuisine, tandis qu'un autre utilisateur u_2 peut d'abord choisir un type de cuisine, puis restreindre les résultats selon le nombre d'étoiles accordées à l'établissement. Une autre qualité que nous pouvons également citer réside dans le fait que l'association de plusieurs facettes d'indexation aux documents offre également l'avantage de créer des liaisons logiques multiples dans un corpus documentaire pour en faciliter le repérage, l'utilisation, et l'exploitabilité des documents sans duplication des unités

documentaires. En 2007, des études et des expériences ont été réalisées au sein de plusieurs organisations publiques, qui ont montré les différents avantages qu'un modèle à facettes présenterait à travers sa souplesse, son expressivité et sa simplicité (Schmetzke *et al.* 2007) (Nasir Uddin et Janecek 2007).

L'exploitation des facettes constitue donc une solution alternative très riche au paradigme hiérarchique traditionnel. Cependant, Ranganathan affirme que l'idée n'est pas de s'en passer du paradigme d'organisation classique, mais de le supporter avec le nouvel aspect multidimensionnel. Les deux paradigmes peuvent être complémentaires, chacun répond à des besoins complémentaires de l'autre (Ranganathan 1931). À ce propos, plusieurs travaux ont pu prouver l'efficacité de cette complémentarité, en instance nous citons les travaux de (Marleau *et al.* 2008) qui proposent un modèle exploitant des facettes et des ontologies sémiotiques pour la gestion documentaire. Aussi, les auteurs (Smith et Shadbolt 2012) qui proposent un modèle nommé « FacetOntology » pour la définition de facettes extraites à partir d'une ontologie. De leur part (da Silva *et al.* 2011) proposent une approche pour améliorer la RI en utilisant l'extension de requêtes par des ontologies alliées à la navigation à facettes. Divers autres travaux ont aussi considéré cette complémentarité (Tvarožek 2006) (Tomasi *et al.* 2015).

En tenant compte de ce petit aperçu, nous présentons dans la figure 2.1 une classification des principales modalités de recherche qui se déclinent en trois grandes catégories. Puis, nous passons dans la section suivante à une revue sur les principaux modèles de représentation et de RI.

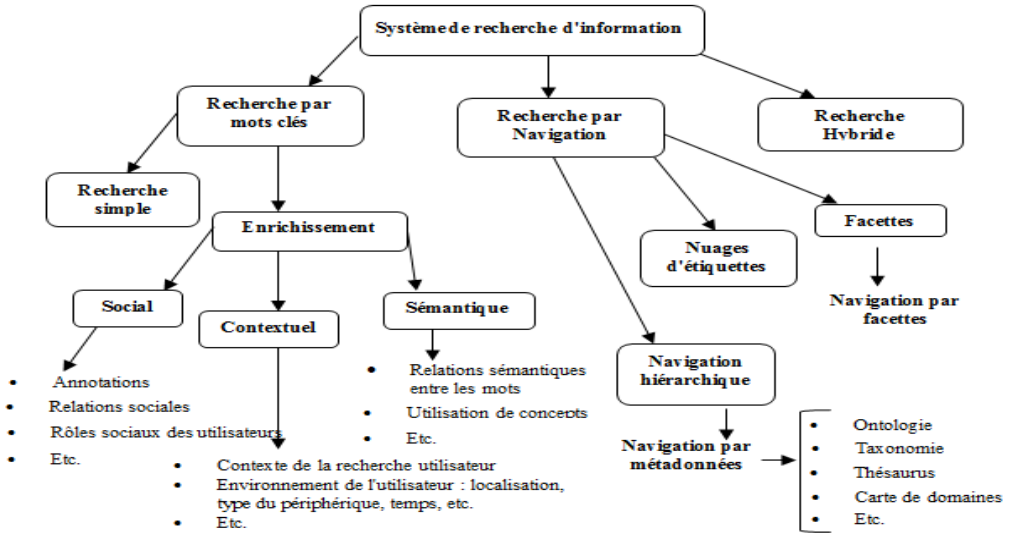


Figure 2. 3. Principales modalités de recherche offerte par un SRI

II.1.4. Modèles de recherche d'information

Afin de chercher les informations qui correspondent à un besoin utilisateur, les SRIs classiques se fondent sur une architecture où l'utilisateur est juste un consommateur du contenu. Nous avons vu que d'une façon générale, un modèle de RI est caractérisé par le modèle de représentation des documents et des requêtes « F » ainsi que du processus d'appariement document-requête « RSV (q, d) ». Il a été défini formellement par un quadruplet (D, Q, F, R (q, d)) (Baeza-Yates et Ribeiro-Neto 1999) où :

- D est l'ensemble des documents de la collection,
- Q est l'ensemble des requêtes des utilisateurs,
- F est le schéma du modèle théorique de représentation des documents et des requêtes,
- RSV (q, d) est la fonction de pertinence du document d à la requête q.

Il existe dans la littérature une grande variété de modèles (Baeza-Yates et Ribeiro-Neto 1999), la figure 3 présente une classification des plus importants, ils sont principalement répartis autour de trois familles : les modèles booléens, vectoriels et probabilistes.

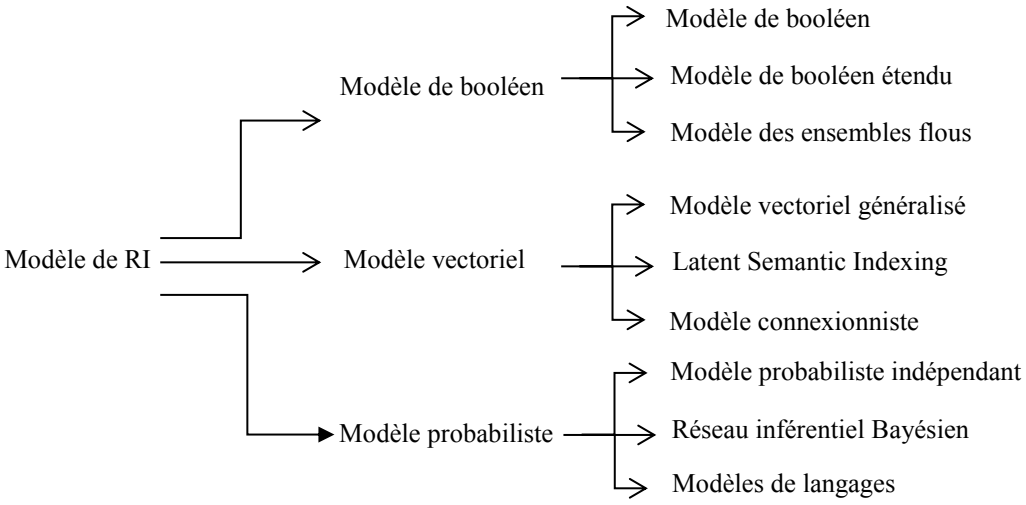


Figure 2. 4. Taxonomie des modèles de RI proposés dans la littérature

II.1.4.1. Modèles booléens

Les modèles booléens, appelés aussi les modèles ensemblistes, sont les premiers des modèles, ils sont basés sur la théorie des ensembles et l'algèbre de Boole (Salton *et al.* 1983). Les documents sont représentés par une conjonction des termes qui constituent leur contenu (par exemple d =

$\{t_1 \Lambda\ t_2 \Lambda, \ldots, \Lambda t_k\}$) et les requêtes sont formulées à l'aide d'expressions logiques (AND, OR, NOT). Un document est jugé pertinent si et seulement si son contenu respecte la formule logique de la requête. Des poids binaires (0 et 1) sont utilisés pour la pondération d'un terme. Ainsi, les termes de la requête sont soit présents, soit absents dans le document. L'inconvénient majeur avec ces modèles en dehors de la difficulté qui se présente lors de la formulation des requêtes, réside dans le fait qu'ils se basent pour la sélection des documents sur la pertinence exacte. Ils considèrent comme non pertinents les documents qui ne contiennent pas tous les termes de la requête utilisateur. Un document est donc soit pertinent, soit non pertinent. De plus, tous les termes ont le même poids lorsqu'ils sont présents dans le document alors qu'en réalité certains sont plus représentatifs que d'autres. Par conséquent, les réponses ne sont pas ordonnées, ce qui ne facilite pas leur examen par le système et leur exploration par l'utilisateur. Ce manque de souplesses peut impliquer un retour considérable de documents lorsque la requête est très large, ou un nombre très réduit dans le cas contraire.

Pour augmenter le niveau de pertinence, le modèle booléen étendu est venu compléter le modèle classique en intégrant des poids d'indexation dans l'expression des documents (Fox 1983) (Salton 1989). Le poids d'un terme est exprimé par le calcul de valeurs de TF (fréquence de terme) et d'IDF (fréquence de document inverse).

- TF: désigne la fréquence du terme dans le document,
- IDF: désigne sa fréquence documentaire inverse, il s'agit de l'importance du terme dans la collection, ainsi les termes qui apparaissent dans peu de documents de la collection ont plus d'importance que deux qui apparaissent dans beaucoup de documents.

Le poids final est obtenu à travers la fonction de pondération TF-IDF suivante (Salton *et al.* 1983):

$$w_{ij} = tf_{i,j} * \log\left(\frac{N}{df_i}\right) \qquad 2.1$$

où

- w_{ij} = *tf*idf*
- $tf_{i,j}$: Nombre d'apparitions du mot dans le document divisé par le nombre d'apparitions du mot le plus fréquent,

- df_i : nombre de documents où apparait le mot,
- N : nombre total de documents du corpus.

Cette extension est une combinaison des modèles booléens et vectoriels. La représentation des documents s'appuie sur les vecteurs de termes pondérés. Ainsi, pour la sélection des documents pertinents correspondant à une requête, une mesure de correspondance algébrique est utilisée, telle que la distance euclidienne. Ceci permet d'avoir un ordonnancement des documents retournés selon leur valeur de pertinence calculée à travers cette mesure de correspondance. Par ailleurs, on trouve les modèles fondés sur la théorie des ensembles flous (Kraft et Buell 1983) qui sont considérés aussi comme une extension de la version classique du modèle booléen. L'intégration de cette théorie dans la RI a permis d'atténuer certaines limitations telles que l'ambigüité des requêtes, l'imprécision dans le processus d'indexation ainsi que l'écart de pertinence entre les documents résultats. L'inconvénient qui a été repéré avec cette théorie est qu'elle ne permet pas d'offrir un ordonnancement à l'ensemble des résultats.

II.1.4.2. Modèles vectoriels

Les modèles vectoriels appelés aussi les modèles algébriques, sont les plus populaires en RI et restent parmi ceux des plus utilisés et des plus efficaces (Russell et Norvig 2010). Les documents et les requêtes sont représentés par des vecteurs de poids dans un espace vectoriel composé de tous les termes d'indexation (Salton et McGill 1986). Cette pondération rend le modèle flexible en permettant un appariement approximatif entre les documents et la requête utilisateur, appelé aussi l'appariement partiel. Il permet d'avoir de meilleures performances comparées aux modèles booléens.

Considérant un document d et une requête q représentés dans un espace de k dimensions respectivement par d = $\{ w_1^d, w_2^d, \dots, w_k^d \}$ et q = $\{ w_1^q, w_2^q, \dots, w_k^q \}$, où w_i^d est le poids du descripteur dans le document et w_i^q est celui dans la requête, la pertinence d'un document vis-à-vis d'une requête peut être définie par des mesures de distances ou de proximités entre les vecteurs associés. Le mécanisme de recherche repose donc sur l'idée de trouver les vecteurs documents qui s'approchent le plus du vecteur requête. Plusieurs mesures ont été proposées dans la littérature, toutes ont l'avantage de donner une liste ordonnée des résultats. La distance la plus connue est la distance euclidienne qui définit l'espace cartésien, nous citons également la mesure du produit scalaire qui est la plus simple à utiliser, la mesure

de corrélation de Pearson, la corrélation Jaccard/Tanimoto, la similarité de Cosinus également souvent utilisée, la mesure de divergence de Bregman, la mesure de Dice , etc. Nous définissons les principales de ces mesures de similarité les plus utilisées dans la littérature :

La mesure de Jaccard :

$$RSV(q,d) = \frac{\sum_{i=1}^{k} w_i^d \, w_i^q}{\sum_{i=1}^{k} (W_i{}^q)^2 + \sum_{i=1}^{k} (W_i{}^d)^2 - \sum_{i=1}^{k} w_i^q * w_i^d} \qquad 2.2$$

Le produit scalaire :

$$RSV(q,d) = \sum_{i=1}^{k} w_i^d \;\; w_i^q \qquad 2.3$$

La mesure Cosinus :

$$\mathrm{RSV(q,d)} = \frac{\sum_{i=1}^{k} w_i^d \, w_i^q}{(\sum_{i=1}^{k} (W_i{}^q)^2)^{1/2} (\sum_{i=1}^{k} (W_i{}^d)^2)^{1/2}} \qquad 2.4$$

La distance euclidienne :

$$RSV(q,d) = \sqrt{\sum_{i=1}^{k} w_i^q - w_i^d} \qquad 2.5$$

Les écrits de Slimani (Slimani *et al.* 2007) , de Choi (Choi *et al.* 2010) et de Pedersen (Pedersen *et al.* 2007) contiennent des explications et des comparaisons très compréhensibles des différentes métriques de similarité et de proximité existantes. Parmi les modèles vectoriels proposés dans la littérature, on trouve également les modèles connexionnistes et en particulier les réseaux de neurones qui imitent les fonctionnalités du cerveau humain (Seidenberg et McClelland 1989). Un réseau de neurones pour la RI peut être composé de trois couches (requêtes, termes, documents). La première couche représente les termes d'une requête, la deuxième représente les termes d'indexation et la troisième représente les documents de la collection. L'interrogation se fait par propagation de signaux de la couche d'entrée vers la couche de sortie via la couche intermédiaire. Autrement dit, la requête utilisateur active la première couche des termes, cette activation est propagée vers la couche des documents, et l'activation finale des

documents donne lieu aux documents résultants du système. Les valeurs de sortie servent de critères de décision reflétant la pertinence des documents, ou l'expansion de requête dans le cas d'une reformulation ou d'enrichissement de son contenu (Kwok 1989) (Kwok 1995).

Ces modèles viennent alléger le langage des requêtes en remplaçant les requêtes booléennes par des requêtes formulées en langage naturel. Toutefois, même si cette démarche présente un avantage pour les utilisateurs, elle ouvre la voie à plusieurs problématiques liées au traitement du contenu de cette requête, celle-ci peut être moins expressive, ou même un peu trop générale par rapport au besoin réel de l'utilisateur, ce qui rend difficile son interprétation par le SRI. Un autre point négatif réside dans la représentation du contenu des documents qui se présente sous forme d'un ensemble de mots, une telle représentation ignore l'ordre des mots et ne prend pas en compte les dépendances entre eux, les mots sont considérés comme des données sans sémantique. Ainsi, les systèmes qui se basent sur de tels modèles se retrouvent rapidement confrontés à de nombreuses limites relatives à l'interprétation de ces mors ayant une influence négative sur la pertinence des résultats offerts. Les facteurs de ces limitations sont abordés avec plus de détails dans la section II de ce chapitre.

Une solution souvent apportée consiste à intégrer une analyse linguistique, il s'agit de ne plus considérer les mots comme de simples mots clés, mais comme des entités linguistiques. Les traitements linguistiques peuvent intervenir à différents niveaux dans le SRI, ils contribuent à créer une représentation plus riche du contenu textuel qui vise à réduire l'écart existant entre les termes utilisés par le système pour l'indexation du contenu et les termes utilisés par les utilisateurs lors de la formulation de requêtes, afin d'obtenir un appariement plus pertinent entre eux. Une ressource linguistique est définie comme étant un ensemble de données spécifiques à une langue particulière et comportant des connaissances linguistiques exploitables par un traitement automatique en particulier (Kayser 1997), nous citons les corpus, les ressources lexicales (ontologies, thésaurus, dictionnaires, etc.) et les grammaires (Cailliau 2010). Cette analyse linguistique regroupe plusieurs disciplines, parmi lesquelles nous citons la sémantique qui étudie le sens du mot (cf. section II.2). Un des modèles qui exploite des ressources sémantiques dans le processus d'indexation est le modèle LSI (Latent Semantic Indexing) (Deerwester *et al.* 1990b), il permet de transformer une représentation à base de mots-clés en une représentation fondée sur les concepts, ceci permet d'un côté de réduire l'espace d'indexation et d'un autre de relier les

documents entre eux. Cette technique d'indexation représente donc une solution au problème d'indépendance de termes repéré avec les autres modèles vectoriels.

Pour plus d'information sur tous les modèles vectoriels nous invitons le lecteur à lire le second chapitre (p.31 à 51) écrit par Romaric Besançon, du livre de référence (Ihadjadene 2004).

II.1.4.3. Modèles probabilistes

Les modèles probabilistes ont également été proposés pour compléter les modèles classiques de la RI, ils s'appuient sur la théorie des probabilités. La pertinence entre un document d et une requête q est mesurée par le rapport entre la probabilité que d soit pertinent pour q, notée p (R/d, q), et la probabilité qu'il soit non pertinent, notée par p ($\bar{R}$/d, q), où R est l'événement de pertinence et $\bar{R}$ de non-pertinence. Ainsi, le modèle de correspondance entre un document d et une requête q, est calculé comme suit :

$$RSV(d,q) = \log \frac{P(R/d,q)}{P(\bar{R}/d,q)} \qquad 2.6$$

Parmi les modèles de RI probabiliste, appelés les dérivés du modèle probabiliste basique, on trouve : le modèle BIR (Binary Independence retrieval) (Yu et Salton 1976) (Robertson et Jones 1976), le modèle inférentiel Bayésien (PEARL 1988), le modèle 2-poisson (Robertson 2005), et le modèle de langage (Ponte et Croft 1998) (Boughanem *et al.* 2004). Comme dans le modèle vectoriel, les documents trouvés dans les modèles probabilistes sont classés par ordre de pertinence par rapport à la requête. Ils sont considérés comme itératifs, l'utilisateur peut intervenir dans le processus pour améliorer les performances. Le modèle BM25 (Robertson et Jones 1976; Robertson et Walker 1994) a été proposé pour pallier les défauts du modèle BIM et reste aujourd'hui l'un des modèles les plus performants observés en RI. Le système Okapi est un exemple d'implémentation de ce modèle (Robertson *et al.* 1996). Ce modèle est un des modèles obtenant de meilleurs résultats en RI classique dans les grandes compétitions (INEX, TREC, etc.). Le poids d'un terme t_j dans document d_i est calculé avec la formule suivante :

$$W_{i,j} = \frac{(k_1+1)*tf_{ij}}{k_1*((b-1)+b*\left(\frac{dl}{avgdl}\right)+tf_{ij}} * \log\left(\frac{N-df_i+0.5}{df_i+0.5}\right) \qquad 2.7$$

Où

- dl : la taille du document d_i et avgdl est la taille moyenne des documents
- tf_{ij} : le nombre d'occurrences du terme t_j dans le document d_i
- k_1: un paramètre qui permet de contrôler la saturation de tf_{ij}
- b: un paramètre qui permet de contrôler la normalisation par rapport à la taille des documents
- N: le nombre de documents dans la collection
- df_i : le nombre de documents qui contiennent le terme t_j

Le score global d'un document pour une requête qui prend en considération le score thématique classique est calculé comme suit :

$$\mathrm{Score}_{\mathrm{BM25}}(q_i, d_j) = \sum_{t_j \epsilon d_j \cap q_i} w_{i,j} \qquad 2.8$$

D'autres extensions de ce modèle ont été proposées, elles prennent en considération pour le calcul de l'importance des mots, d'autres facteurs, en plus du score thématique classique, notamment, le modèle BM25t qui considère l'importance des balises dans le cas des documents XML (Géry *et al.* 2010), et le modèle BM25F qui considère l'importance des champs de description dans le cas des documents structurés en champs (Robertson *et al.* 2004), et le BM25FS qui considère l'apport du contexte social dans le cas de la recherche sociale (Bouhini *et al.* 2013a).

II.1.5 Évaluation des systèmes de recherche d'information

La validation d'un nouveau SRI se base sur l'évaluation expérimentale de ses performances. Au cours des dernières années, cette évaluation a été un domaine de recherche très actif, elle permet d'estimer l'impact de chacune des caractéristiques d'un système et de fournir des paramètres objectifs de comparaison entre les différents systèmes existants. L'évaluation peut couvrir plusieurs critères d'efficience et d'efficacité qui sont en général construits à partir des jugements exprimés par des utilisateurs ou par des experts. On peut citer : la pertinence des résultats, la qualité de la présentation des résultats, la performance qui touche à son tour plusieurs critères concernant la consommation de ressources, telle que le temps de réponse, l'espace mémoire, la capacité en charge, etc. Le critère le plus important qui sans doute intéresse le plus l'utilisateur est celui qui mesure la capacité qu'un SRI puisse satisfaire son besoin en information. Il s'agit de la pertinence des résultats renvoyés par ce système.

Comme nous l'avons vu dans les sections précédentes, la RI est centrée sur cette notion de pertinence qui définit le degré de correspondance entre une requête de recherche et les documents dans l'index. En réponse à cette requête, la liste des résultats peut être divisée en quatre ensembles: pertinents ou non (l'utilisateur dit Oui / Non), sélectionnés ou non (le système dit Oui / Non). Ces différentes situations sont résumées dans le tableau suivant, appelé la table de contingence.

	Pertinent	Non pertinent
Sélectionnés	Vrai positif	Faux positives
Non sélectionnés	Faux négatives	Vraies négatives

Tableau 2. 1. Table de contingence

À partir de ces valeurs, différentes métriques peuvent être calculées, permettant d'évaluer la pertinence des résultats, nous citons les deux métriques les plus utilisées en RI, à savoir, le taux de précision qui tente de répondre à la question: combien d'éléments sont pertinents parmi ceux qui sont sélectionnés, il mesure la capacité du système à rejeter tous les documents non pertinents, tandis que le taux de rappel tente de répondre à la question: combien d'éléments pertinents sont sélectionnés, il mesure la capacité du système à sélectionner tous les documents pertinents.

$$Précision(P) = \frac{vrai\ positifs}{vrai\ positif + faux\ positifs} = \frac{Nombre\ de\ documents\ pertinents\ selectionnés}{Nombre\ de\ documents\ selectionnés} \qquad 2.\ 9$$

$$Rappel(R) = \frac{vrai\ positifs}{vrai\ positif + faux\ négatifs} = \frac{Nombre\ de\ documents\ pertinents\ selectionnés}{Nombre\ de\ tous\ les\ documents\ pertients} \qquad 2.\ 10$$

Le rappel n'est pas une mesure suffisante, car il ne montre pas qu'il y a des documents non pertinents parmi les résultats. Ces deux métriques sont dépendantes l'une de l'autre, quand l'une augmente, l'autre diminue. Autrement dit, plus le bruit[3] est grand, plus la précision est faible, et vice versa. Différents travaux ont montré que l'utilisation seule de ces deux mesures est insatisfaisante, cette critique porte sur l'inadéquation de l'évaluation binaire sur laquelle sont basées ces mesures (pertinent et non pertinent), les auteurs pensent que certains documents retrouvés peuvent être partiellement pertinents

[3] Nombre de résultats non pertinents

(Harter et Hert 1997). Des travaux ont tenté d'améliorer ces mesures et d'autres ont proposé de nouvelles, à noter la mesure ESL (Expected search length) proposée par Cooper, qui correspond au nombre de documents non pertinents que l'utilisateur doit parcourir dans la liste des résultats avant d'accéder à un document pertinent. Dans le même ordre d'idée une autre mesure basée rang a été proposée, elle est souvent utilisée dans les systèmes questions-réponses, il s'agit de la métrique MRR (Mean Reciprocal Rank). Elle permet d'évaluer la moyenne du rang du premier document pertinent dans la liste des résultats, calculée sur l'ensemble des requêtes. Elle est définie comme suit :

$$MRR = \frac{1}{|Q|}\sum_{i=1}^{Q}\frac{1}{rang_i} \qquad 2.11$$

La valeur de MRR est élevée pour un système qui donne des résultats pertinents en haut de la liste renvoyée, et elle est nulle si aucun document pertinent n'a été retrouvé.

On trouve également les mesures à X documents et la précision moyenne. La précision à X documents notée par P@X (où X peut prendre différentes valeurs : 5, 10, 15, etc.), représente le nombre de documents pertinents P_t sur les X premiers.

$$P@X = \frac{P_t}{X} \qquad 2.12$$

Cette pertinence est souvent reliée à ce que l'on nomme la précision exacte ou la R-précision. Si n documents pertinents existent dans le corpus pour une requête de recherche q, la précision exacte est celle calculée pour les n premiers documents de la liste ordonnée des résultats retournés suite à cette requête. La précision moyenne quant à elle utilise uniquement les valeurs de précision où un document de la liste ordonnée est pertinent, et calcule leur moyenne sur l'ensemble des requêtes.

$$MAP = \frac{\sum_{q\epsilon Q} AP_q}{|Q|} \qquad 2.13$$

Avec AP_q la précision moyenne pour une requête, Q l'ensemble des requêtes, et $|Q|$ le nombre de requêtes dans l'ensemble. Cette mesure tient compte à la fois de la précision et du rappel. Elle peut être calculée à différents niveaux du rappel, pour chaque niveau les valeurs calculées sont aussi moyennées sur l'ensemble des requêtes.

Nous citons également la F–mesure qui combine la précision et le rappel, nommée aussi F-score, elle définit par la formule suivante :

$$F_\beta = \frac{(1+\beta^2)*(precision*rappel)}{(\beta^2*precision*rappel)} \quad 2.\ 14$$

Où β traduit l'importance relative au rappel et à la précision, un cas particulier est la mesure F1 avec β =1 qui définit la moyenne harmonique de ces deux mesures comme suit

$$F_1 = \frac{2*(precision*rappel)}{precision*rappel} \quad 2.\ 15$$

II.1.5.1 Protocoles d'évaluation

Les premiers protocoles adoptés en RI sont initiés par Cleverdon (Cleverdon 1967) dans le cadre du projet Cranfield. Ils se basent sur une approche de type laboratoire, dite en anglais « laboratory-based model ». Elle constitue le cadre de référence dans lequel s'inscrivent les expérimentations et la validation des systèmes classiques. Cette approche fournit des ressources de test composées en général d'un triplet d'éléments : un ensemble de documents, un ensemble de requêtes et un ensemble de jugements de pertinence associés à chaque requête de la collection. Ce type de collection est souvent adopté dans les compagnes d'évaluation des SRI telles que TREC (Text REtrieval Conference), CLEF (Conference and Labs of the Evaluation Forum), INEX (Initiative for the Evaluation of XML Retrieval), etc. Les collections TREC représentent aujourd'hui un référentiel incontournable en RI, ils comportent un corpus de données et des évaluations binaires afin d'évaluer les performances des SRIs. TREC n'est plus juste une collection, mais devenu depuis un bon moment un programme d'évaluation des SRIs, initié par le l'Institut national des normes et de la technologie (NIST) aux États-Unis. Il propose une plate-forme pour l'évaluation et la comparaison d'expérimentations sur des collections volumineuses de textes, elle comporte des collections de tests, des tâches spécifiques et des protocoles d'évaluation pour chaque tâche.

II.1.5.1.1. Description d'une tâche

Le principe général d'une tâche est que l'on dispose d'une collection de requêtes qui représente des besoins en information, d'une collection de documents et d'un ensemble complet de valeurs de pertinence qui indique si l'association requête-document a été jugée satisfaisante ou invalide selon l'appréciation des évaluateurs (experts ou utilisateurs). Nous citons par exemple la tâche Ad-hoc dans TREC qui évalue les performances des SRIs sur des ensembles statiques de documents avec des requêtes dynamiques qui changent. Cette tâche peut être considérée comme étant similaire à une recherche dans une bibliothèque, où la collection est connue contrairement aux requêtes de recherche qui ne le sont pas. Le principe est de créer d'abord des requêtes à partir des besoins en information soumis par de vrais utilisateurs. Pour l'évaluation, chaque participant fournit au NIST la liste des 1000 premiers documents retrouvés par son système en réponse à chacune des requêtes. Les évaluateurs jugent la pertinence des 100 à 200 premiers documents de chaque système puis différentes mesures d'évaluation sont appliquées, à savoir le rappel et la précision, la précision moyenne, la précision à X premiers documents, etc.

II.1.5.1.2. Collection de tests

La collection de tests (dite aussi corpus de tests) sur laquelle se base l'évaluation orientée-laboratoire, constitue le contexte d'évaluation des systèmes. Elle comprend les éléments qui vont servir à tester le processus de sélection des documents suite à une requête de recherche. Les collections diffèrent par plusieurs facteurs : i) par le nombre de documents et de requêtes qui les constituent, ii) par domaine ou spécialité auquel ces éléments appartiennent, et aussi ii) sur la façon de juger la pertinence pour associer les documents aux requêtes (évaluation basée sur des experts ou sur de vrais utilisateurs), etc. La méthodologie Cranfield utilise une collection de documents, un ensemble de sujets utilisé pour définir les requêtes, et les jugements de pertinence correspondants, avec des exemples de documents pertinents et non pertinents pour les requêtes générées à partir de sujets. Les collections offertes sont de l'ordre de quelques giga-octets et de quelques centaines de giga-octets pour les grandes collections.

Collection de documents : c'est l'ensemble de documents sur lesquels les SRIs emploient des requêtes de recherche et sélectionnent les éléments pertinents. Ces documents peuvent provenir de différentes sources de données dont généralement la presse écrite, telle que le Wall Street Journal, mais

également des documents web tels que des résumés de cours de publications, des documents informatiques, des brevets, etc. Ils peuvent aussi être structurés par des annotations indiquant la présence d'un titre, le renvoi aux paragraphes, etc. Ces documents diffèrent par plusieurs critères, à noter la longueur, le genre, la langue, le format (SGML ou DTD), et la date. Le choix de la collection en termes de contenu et de longueur est effectué selon la tâche de recherche que le système veut évaluer, cela permet de garantir une représentativité par rapport à cette tâche et garantir une diversité des sujets et de vocabulaire.

Collection de requêtes : ce corpus simule les activités de recherche de l'utilisateur à travers des sujets (topics) formulés sous forme de textes à partir desquels les requêtes sont définies. Ces requêtes sont formulées soit par des assesseurs de la compagne d'évaluation ou extraits à partir des fichiers log de recherche où les requêtes sont employées par de vrais utilisateurs. Le corpus TREC utilise de 25 à 50 sujets comme une norme de test en termes de nombre de requêtes. Elles sont adéquates en longueur, en thèmes abordés et par forme. Ce corpus suit un modèle de base qui définit pour chaque requête un titre (sujet), un numéro, et une description qui indique les documents qui doivent être pertinents et également ceux qui ne doivent pas l'être. Ce modèle est illustré par l'exemple suivant :

```
<top>
<head> Tipster Topic Description
<num> Number: 062
<dom> Domain:  Military
<title> Topic:  Military Coups D'etat
<desc> Description: Document will report a military coup d'etat,
either attempted or successful, in any country.
<smry> Summary: Document will report a military coup d'etat,
either attempted or successful, in any country.
</top>
```

Figure 2. 5. Modèle de base de TREC pour le corpus de requêtes

Le jugement de pertinence : cette pertinence constitue la tâche la plus difficile. Un degré de pertinence (binaire ou numérique sur une échelle de 0 à 2) est attribué pour chaque document par rapport à chaque requête. Pour juger cette pertinence, des évaluateurs prennent en charge cette tâche en examinant le contenu de chaque document du corpus documentaire et évaluent sa pertinence pour chaque requête. Dans le cas de grandes collections telles que le corpus TREC, la technique de pooling[4] est

[4] Le pooling est une technique permettant de constituer un ensemble varié de documents à partir de plusieurs modèles, puis trouver une pertinence commune sur ces documents en se basant uniquement sur le critère de sujet ou la notion de pertinence thématique.

utilisée pour établir les jugements de pertinence. Elle est effectuée à partir des 1000 premiers documents retrouvés par les systèmes participants. L'idée est d'examiner pour chaque requête un ensemble des documents identifiés par différents moteurs de recherche, au lieu d'un seul document. L'ensemble de documents est obtenu par l'exécution de dizaines d'algorithmes de recherche (Buckley et Voorhees, 2004), il diffère d'une requête à l'autre, et varie entre 1000 et 2000 documents. Le rappel calculé est relatif à l'ensemble des documents examinés pour chaque requête. Le résultat est un fichier nommé « Qrels » regroupant les jugements, dont la structure est comme suit : SUJET ITERATION DOCUMENT# PERTINENCE, sachant que la pertinence est sous forme d'un score de pertinence qui peut être binaire ou numérique sur une échelle de 0 à 2 pour indiquer une grande pertinence.

II.2. Partie 2 : Émergence de la recherche d'information adaptative

II.2.1. Facteurs d'émergence

Plusieurs limitations ont poussé les chercheurs à adapter la RI classique, parmi lesquelles nous pouvons citer :

Faiblesse dans la représentation de l'information et dans la correspondance « document-requête »: une problématique cruciale avec la RI classique est l'écart considérable qui peut exister entre les univers de représentation utilisés pour interpréter d'une part le besoin informationnel des utilisateurs, et d'un autre, la collection de documents disponible pour la recherche. Ce problème est dû principalement à l'utilisation des mots clés pour la représentation des documents, qui n'informe nullement sur les relations qui peuvent exister entre les termes d'indexation, ce qui engendre une dégradation de performance dans le processus de recherche. Cette dégradation est due aussi à l'insuffisance de l'appariement sur lequel se base la recherche classique pour sélectionner les documents, cette recherche se base uniquement sur la ressemblance exacte ou lexicale entre les mots, il en résulte ainsi un défaut d'appariement. Plus concrètement, deux limitations principales peuvent être soulevées, nous citons i) l'absence de relations sémantiques entre les mots qui réduit l'accès à l'information pertinente. En effet, une même idée peut être exprimée de différentes manières. La principale conséquence de cela est de ne pas pouvoir retourner à l'utilisateur un document qui correspond à des termes sémantiquement proches à sa requête, mais

toutefois différents, tels que des synonymes, des hyperonymes ou hyponymes. Ainsi, une requête contenant par exemple le terme « résidence » ne pourra pas retrouver un document contenant le mot « maison » ou « logement ». Aussi une page web d'un tutoriel Java ne pourra jamais être sélectionnée en réponse à une requête sur « les langages de programmation » si les mots « langages de programmation » sont absents de cette page. Deuxièmement, nous citons l'absence de traitement du phénomène traditionnel de la langue naturelle, le phénomène de polysémie, qui se manifeste lorsqu'un mot dans la requête ou un terme d'indexation a plusieurs significations, ceci peut induire à des résultats non pertinents. En conséquence, le terme « virus » présent dans la requête de l'utilisateur peut à la fois renvoyer à des documents parlant d'informatique ou de santé. Cette ambigüité peut conduire donc à une diminution de performance du système. Ainsi, nous pouvons dire que les mécanismes de représentation du contenu et d'appariement document-requête ont un impact direct sur la qualité des résultats fournis.

Le manque d'expertise de l'utilisateur et la faible expression de son besoin informationnel à travers ses requêtes de recherche: une des tentatives les plus utilisées est la considération de l'aspect sémantique pour améliorer la pertinence des résultats de recherche. Cela consiste en l'identification de relations sémantiques qui peuvent exister entre les termes d'indexation ou entre les termes de la requête et ceux des documents (Rosso *et al.* 2004). Néanmoins, cette pertinence ne dépend pas seulement de ces termes d'indexation ou du mécanisme d'appariement qui peut être appliqué lors de la recherche, mais aussi de façon non négligeable de l'utilisateur qui n'exprime pas toujours correctement ses besoins en information. Ainsi, les requêtes sont souvent courtes et moins expressives, en particulier lorsque cet utilisateur n'a pas assez de connaissances sur le domaine de sa recherche, ou simplement a du mal à traduire ses besoins sous la forme de mots clés. Ceci est connu sous le problème de l'inadéquation des besoins réels utilisateur avec sa requête. Les mots clés exprimés par l'utilisateur peuvent aussi être propres à lui, c'est-à-dire, ils sont personnels et ne correspondant pas forcément aux mots utilisés par le système pour décrire les documents, ceci est connu sous le problème de la non-concordance entre les espaces de représentation documents-requêtes. Les termes choisis ont donc une grande influence sur les réponses qui peuvent être retournées par le système. On remarque que souvent l'utilisateur relance ses recherches lorsqu'il est n'est pas satisfait des résultats, en reformulant sa requête ou en augmentant son contenu avec d'autres termes dans le but de combler un manque d'informations. Ces informations

peuvent être liées à plusieurs catégories d'information: à une ressource d'enrichissement, à des préférences ou à des centres d'intérêt de l'utilisateur, à l'environnement où l'utilisateur effectue sa recherche, ou à une activité ponctuelle, c'est-à-dire, le moment où la recherche est effectuée (Belkin et Croft 1992), ou autre.

La non-reconnaissance de l'utilisateur par le système et l'absence de son contexte de recherche: par-dessus toutes ces contraintes citées, les systèmes classiques ne permettent pas de reconnaitre les utilisateurs donc ils retournent le même ensemble de résultats pour la même requête envoyée par différents utilisateurs sans tenir compte de leurs critères spécifiques ou du contexte de leur recherche. Un des problèmes immédiats posés par une telle approche est notamment l'ambiguïté du sens des mots. Prenons l'exemple de « Java magazine », cette requête réfère à plusieurs thèmes, elle peut référer à un langage de programmation ou à une ile déserte ou à un magazine ayant le nom de Java. Aussi, la requête « Resident evil » qui se rapporte à plusieurs interprétations (série de jeux de vidéo, série de films cinématographiques). L'utilisation des moteurs de recherche classiques telle que Google et Bing, oblige l'utilisateur à préciser sa recherche pour recevoir du contenu pertinent qui répond à ses attentes. D'autres facteurs viennent intensifier ce problème de non-pertinence des résultats, notamment, la quantité considérable d'information disponible pour la recherche, et les expressions plus ou moins larges du besoin en information des utilisateurs, etc. (Xu 1997). Par exemple, la requête « restaurant gastronomique » est une requête locale qui doit renvoyer des résultats concernant les restaurants de la région ou de la ville dans laquelle est localisé l'utilisateur. En occurrence, cette requête doit renvoyer à Montréal les restaurants gastronomiques dans cette région. Ces facteurs constituent le contexte de recherche de l'utilisateur et influencent négativement sur la pertinence des résultats fournis.

II.2.2. Dimensions d'adaptation

Plusieurs approches sont proposées dans le but d'améliorer les performances de ces SRIs, notamment les techniques qui visent à adapter le processus de recherche au besoin précis de l'utilisateur. Le processus d'adaptation s'effectue en incluant des éléments additionnels en dehors des principaux éléments du SRI classique. Cette adaptation peut être effectuée à différents niveaux du processus RI. On trouve i) l'adaptation du contenu informationnel des documents, ii) l'adaptation du besoin informationnel

de l'utilisateur, iii) l'adaptation de l'accès à l'information et v) l'adaptation de la présentation des résultats.-Ces techniques font l'objet de la RI adaptative et se résument comme suit :

- ✓ Lors de la représentation du contenu documentaire, le système utilise des métadonnées sur les documents ou d'autres catégories d'enrichissement pour une meilleure représentation des documents.
- ✓ Lors de l'interprétation de la requête, le système intègre des informations additionnelles pour mieux cibler le besoin informationnel effectif de l'utilisateur.
- ✓ Lors du processus de recherche, le système inclut des informations pour calculer la pertinence d'un document.
- ✓ Lors de l'affichage des résultats, le système retourne les résultats puis les affiche selon un paramètre défini par le système (ex. les préférences de l'utilisateur, une base de connaissances externe telle que les ontologies et les taxonomies, etc.).

II.2.2.1. Adaptation du contenu informationnel des documents

L'adaptation du contenu documentaire consiste à l'enrichir et renforcer sa représentation avec des informations additionnelles qui aident à faciliter son exploitation et accélérer son accès en recherche. Ces informations, appelées aussi les métadonnées, et peuvent être de différentes catégories : sémantiques, contextuelles, ou sociales.

II.2.2.1.1. Annotation sémantique

Parmi les diverses techniques d'adaptation du contenu documentaire, nous distinguons celles basées sur l'enrichissement sémantique au niveau de l'indexation. Nous citons à titre d'exemple l'utilisation des concepts qui peuvent être implicites, c'est-à-dire, découverts par des calculs statistiques sur la collection de documents, ou explicites extraits d'une ressource sémantique externe. On parle dans les deux cas de l'indexation conceptuelle qui consiste à représenter des documents au moyen d'entités sémantiques (les concepts) plutôt que les entités lexicales (les mots-clés). Ceci permet principalement de rapprocher les informations entre elles et de réduire l'espace d'indexation traditionnel. L'idée des concepts implicites consiste à construire ceux qui couvrent le contenu des documents en se basant sur les relations qui relient les termes dans ces documents (Koll 1979). Cette relation peut être de différentes natures (relation sémantique, relation de cooccurrence (Deerwester *et al.* 1990a), etc.). L'avantage d'une telle approche est

qu'aucune ressource de connaissance externe n'est requise, ce qui la rende adaptable à toute collection de documents. Toutefois, une étape d'analyse du contenu est nécessaire, qui peut être très coûteuse en termes de temps et d'effort. Aussi, la gestion des mises à jour des documents (ajout, modification, suppression des documents) peut requérir la mise à jour de l'espace des concepts qui les représentent.

D'un autre côté, on trouve l'enrichissement basé sur les ressources sémantiques externes telles que les ontologies et les taxonomies de concepts. Ceci a fait apparaitre ce qui est connu par l'indexation contrôlée. Dans un tel enrichissement, l'index est construit sur la base des concepts présents dans une ressource prédéfinie et non sur base des mots qui constituent le contenu des documents. Ce mode d'indexation a donné naissance à plusieurs ressources sémantiques, notamment à i) un modèle formel conçu en 1964 similaire à ce qui est connu aujourd'hui par l'ontologie, au ii) thesaurus Wordnet, devenu aujourd'hui la ressource la plus utilisée dans la RI compte tenu de sa simplicité et de sa richesse dans la langue anglaise, etc. Plusieurs systèmes ont utilisés WordNet pour la conceptualisation des documents (Desmontils et Jacquin 2001) (Desmontils *et al.* 2002) (Baziz 2005) (Boubekeur *et al.* 2010) fusionné à d'autres ressources sémantiques, telle que l'ontologie Penman (Guarino *et al.* 1999). Cependant, avec l'utilisation de WordNet une phase de désambiguïsation est nécessaire, puisqu'un terme peut être lié à plusieurs significations dites aussi synsets (Stetina et Nagao 1998). Cette conceptualisation se déroule généralement en trois étapes: i) l'extraction des mots-clés, ii) l'identification de leur sens (concepts) et ii) la pondération des concepts. De manière analogue à la pondération des termes dans la RI classique basé sur les mots clés, les pondérations des concepts ont pour but d'attribuer à chaque concept son importance dans un document.

Par ailleurs, on distingue l'indexation sémantique qui prend en compte la sémantique des mots à travers des relations entre les termes d'indexation. D'après (Baziz 2005) cette indexation est basée sur l'attribution des significations aux mots en s'appuyant sur des techniques de désambiguïsation de mots (WSD), ainsi l'entrée de l'index est de type « descripteur-signification » plutôt que de simples descripteurs. Une manière d'indexer serait par exemple d'associer aux descripteurs des informations du contexte qui aident à déterminer leur signification : par exemple, Java/informatique et Apple/fruit (Yarowsky 1993) (Sanderson 1994). Nous citons également les travaux de (Voorhees 1993), (Krovetz et Croft 1989), (Sanderson 2000), (Gonzalo *et al.* 1998), (Uzuner *et al.* 1999) qui ont exploité le sens des mots dans la RI. Des méthodes similaires combinent les contextes locaux et globaux des mots pour

capturer leur sémantique dans les documents (Huang *et al.* 2012). Le contexte local d'un mot est son voisinage, c'est-à-dire, l'ensemble des mots qui l'entourent, dits aussi son entourage, tandis que son contexte global du mot est le document entier auquel il appartient. D'autres techniques de désambiguïsation plus élaborées se basent sur la similarité ou la distance sémantique entre les mots à comparer. Cette notion de correspondance sémantique peut être aussi utilisée dans l'indexation conceptuelle pour comparer entre les concepts d'une ontologie. À cet effet, plusieurs mesures ont été proposées dans la littérature permettant la comparaison entre deux termes ou deux concepts, on trouve les approches qui se basent sur les arcs pour l'évaluation de la distance qui sépare les objets dans une ontologie, tels que la mesure de We et Palmer (Wu et Palmer 1994), la mesure de Rada (Rada *et al.* 1989) et celle de Ehrig (Ehrig *et al.* 2005). Aussi les approches basées sur les nœuds qui adoptent une nouvelle mesure entropique de la théorie de l'information, on distingue alors la mesure de Resnik (Resnik 1995), la mesure de Lin (Lin 1998), et celle de Hirst et Onge (Hirst et St-Onge 1998), etc. Puis les approches hybrides qui combinent entre les deux techniques de similarité précitées.

Le problème majeur avec les mesures exploitant les ontologies de référence est que dans certaines situations, la similarité entre deux nœuds voisins dépasse celle de deux nœuds contenus dans la même hiérarchie. Afin de pallier cette limitation, des auteurs proposent une mesure d'extension de la mesure de Wu et Palmer, cette extension se base sur une fonction pénalisant la similarité de deux nœuds éloignés n'appartenant pas à une même hiérarchie (Slimani *et al.* 2007). D'autres parts, on distingue aussi les approches basées sur l'espace vectoriel que nous avons déjà vu dans la section II.1.4.2, tels le cosinus, la similarité euclidienne, la similarité de Jaccard, etc.

Les méthodes sémantiques sont liées le plus souvent chacune à un domaine spécifique. En instance, nous citons les travaux (Lacoste *et al.* 2006) (Maisonnasse *et al.* 2008) (Maisonnasse *et al.* 2009) qui ont utilisé pour indexer une collection d'images, le système de langage médical unifié UMLS. Par ailleurs, dans (Zhou *et al.* 2007) les auteurs ont utilisé le thésaurus médical MeSH (Pollitt 1988), pour indexer les documents TREC (Text REtrieval Conference). Plusieurs d'autres approches ont été proposées dans la littérature qui sont aussi fondées sur des ontologies du domaine, à noter le domaine médical (Pouliquen

2002) (Abdulahhad *et al.* 2011), juridique (Berrueta *et al.* 2006), biopuces[5] (Khelif et Dieng-Kuntz 2004), artistique (Petersen 1994), de transports publics (Maedche et Staab 2001), etc. D'autres ressources sémantiques ont également été utilisées pour l'indexation dans des contextes plus génériques. Nous pouvons citer l'approche "Key Concepts" proposée par (Gauch *et al.* 2003b) qui se base sur les concepts d'ODP (Open Directory Project), et celle de (Labrou et Finin 1999) qui utilise la hiérarchie générique de concepts Yahoo pour classer les documents dans un index hiérarchique.

II.2.2.1.2. Annotation contextuelle

Contrairement à l'annotation sémantique, l'annotation contextuelle ne dépend pas seulement du contenu des documents, mais aussi de leur contexte. Deux informations contextuelles exploitées pour annoter les documents sont distinguées, la première utilise les liens de référencement entre ces documents et la deuxième utilise les liens de citations entre eux. Les relations de citations entre documents sont beaucoup plus utilisées dans les réseaux bibliographiques et scientifiques, elles se traduisent par le référencement des documents dans le contenu des autres. L'analyse de cette relation permet par exemple de déterminer l'impact d'un auteur dans un domaine particulier, en évaluant le nombre de fois où cet auteur a été cité. On distingue aussi la relation de co-citation utilisée en bibliométrie depuis 1973, qui indique la fréquence avec laquelle deux documents sont cités ensemble par les mêmes documents. Ces relations sont exploitées pour évaluer une similarité thématique entre les documents. L'hypothèse sur laquelle se base ce principe est que deux documents qui citent un ou plusieurs autres documents communs ont une relation significative qui se traduit par le nombre de citations en commun (Kessler 1965). Les documents citant sont donc couplés biblio graphiquement s'ils partagent au moins une référence bibliographique. Cette méthode de couplage a été vite critiquée par (Kessler 1963) qui spécifie qu'un document scientifique peut traiter plusieurs sujets, et peut être cité pour plusieurs thèmes différents, cette similarité basée sur la relation de couplage peut donc être inefficace, en particulier lorsque la force de couplage est faible. Autrement dit, deux documents cités par le même document ne peuvent pas être toujours considérés comme similaires, car ils peuvent être cités pour deux thèmes différents, tandis que quand deux documents citent plusieurs mêmes autres documents, la probabilité pour que tous les documents soient cités pour un thème différent est faible.

[5] Permettent de dépister, de détecter, d'identifier des séquences d'ADN grâce aux propriétés des acides nucléiques.

Un autre type d'annotation est le lien de référencement entre les documents qui se traduit par un lien de citation sous la forme d'un hypertexte. En parlant de liens référentiels, on se réfère au système hypertexte, qui se présente sous un graphe orienté, les nœuds représentent les pages web et les arcs sont les liens hypertextes qui les relient. Ce type d'annotation a été exploité pour évaluer l'importance des pages web, à ce propos nous citons les deux algorithmes les plus connus de classement de pages : l'algorithme Page Rank utilisé par le moteur de recherche Google (Brin et Page 1998), et l'algorithme Hits proposé par Jon Kleinberg et considéré comme le précurseur l'algorithme Page Rank (Kleinberg 1999). Le principe du premier algorithme Page Rank est d'évaluer l'importance d'une page en fonction des pages qui pointent vers son contenu. L'algorithme hits quant à lui permet de mesurer l'autorité d'une page web par rapport à d'autres. On dit qu'une page est autoritaire lorsqu'elle a beaucoup d'informations pertinentes et moins de liens hypertextes par rapport à d'autres. On peut aussi citer les autres travaux exploitant les liens pour d'autres objectifs comme la découverte de communautés dans les réseaux sociaux (Gibson *et al.* 1998), (Kumar *et al.* 1999) (Vandaele *et al.* 2004) (Clauset 2005), la classification de pages basée sur le calcul de similarité entre les liens hypertextes (Phelan et Kushmerick 2002), la portée géographique d'un document (Buyukokkten *et al.* 1999) (Thilliez et Delot 2004), etc. Le contexte de référencement inclut aussi les méthodes de propagation de mots clés sur les pages web (Marchiori 1998), de métadonnées sur le Web (Prime-Claverie 2004) et de signatures lexicales entre pages web (Bouklit et Lafourcade 2006).

D'autres dimensions contextuelles sont identifiées et utilisées dans la littérature pour représenter le document, celles-ci ne dépendent pas directement de son contenu textuel. Elles concernent les caractéristiques de la source des documents (Xie 2008) telles que sa crédibilité et sa fiabilité. Nous citons également la qualité de l'information, dans ce cas la notion de pertinence est étendue à une notion liée à plusieurs paramètres tels que la fraîcheur, la sécurité, la précision, la cohérence, etc.

II.2.2.1.3. Annotation personnalisée: annotation sociale et folksonomies

Avec l'avènement du web 2.0, les technologies de l'indexation sociale prennent le dessus dans l'organisation, le classement, et l'accès aux informations. Cet aspect social vient renforcer l'indexation traditionnelle et sémantique par l'association des mots clés libres, au contenu des documents, appelés les

étiquettes ou les tags. Cette pratique a gagné beaucoup en popularité sur le Web et a permis de construire ce qui est connu par les systèmes folksonomiques, dont des exemples les plus populaires sont Delicious[6] et Flickr[7]. Dans Delicious, les utilisateurs annotent des documents, d'autre part, Flickr permet aux utilisateurs de télécharger, de partager et de gérer des images. D'autres applications sont spécialisées dans la musique, les blogs ou les publications de journaux.

L'ensemble des étiquettes individuelles forme une collection qui s'appelle « personomy », et l'ensemble de ces personomies constitue la « folksonomy » (Hotho *et al.* 2006). Ce concept de folksonomie a été introduit par (Vander 2007), il représente un mélange entre deux termes anglais exprimant une idée de classification « taxonomy », faites par des gens « folks », il est né du besoin de décrire la manière dont les utilisateurs emploient des mots clés de façon collaborative pour décrire et organiser l'information. Selon (Mika 2005), une folksonomie désigne un système de classification collaborative basé sur une indexation spontanée, effectuée par des utilisateurs non spécialistes. Formellement, cette folksonomie est composée de trois entités: un ensemble d'utilisateurs, un ensemble d'étiquettes et un ensemble de ressources qui peuvent être de différentes catégories : sites web, livres, vidéos, photos, etc. L'utilisateur est l'acteur principal du système et contribue à la production du contenu par l'ajout de ressources et l'affectation d'étiquettes. Ces descripteurs offrent à l'information de nombreux chemins d'accès qui contribuent à l'amélioration de sa recherche (Hassan-Montero et Herrero-Solana 2006; Jomsri *et al.* 2009).

Cette pratique d'étiquetage a été ensuite étendue pour inclure ce que l'on appelle les nuages d'étiquettes, il s'agit de groupes d'étiquettes provenant de différents utilisateurs, qui rassemblent des informations sur la popularité de ces annotations. Ces informations sont souvent affichées sous la forme d'un nuage dans lequel les étiquettes ayant une grande fréquence d'utilisation sont mises en avant dans un texte plus grand (Hassan-Montero et Herrero-Solana 2006). Cette extension a rendu possible la recherche à base de nuages d'étiquettes, l'utilisateur n'a donc pas besoin de réfléchir aux mots-clés employés pour ses recherches, mais seulement de cliquer sur les étiquettes qui peuvent répondent à ses attentes. Cette technique a permis de compléter les modes actuels d'organisation et de navigation qui sont souvent sous

[6] Delicious: www.delicious.fr
[7] Twitter: www.Flickr.fr

forme de cases à cocher ou de menus déroulants, et de faciliter l'accès aux ressources par une description visuelle de leur contenu, via des annotations, ce qui permet de personnaliser leur accès selon les besoins en information des utilisateurs (Hassan-Montero et Herrero-Solana 2006).

Plusieurs approches ont intégré l'aspect social dans la RI, ces approches peuvent être divisées en deux catégories: la première classe représente les approches qui se basent sur le contenu informationnel apporté par l'utilisateur, à savoir, ses traces d'annotations en termes de ressources et d'étiquettes, la deuxième classe quant à elle intègre en plus de ces traces d'annotation, les relations sociales entre les utilisateurs.

Nous commençons par citer les travaux de (Zhou *et al.* 2008) qui proposent un modèle de génération d'annotations permettant de déduire les thèmes associés aux documents en supposant qu'une étiquette et un sujet du document peuvent refléter la même chose. De leurs parts (Bao *et al.* 2007) et (Xu *et al.* 2007) exploitent les étiquettes pour évaluer la similarité entre les documents et montrent qu'une telle exploitation permet d'optimiser la RI sur le web. Pour ce faire, un algorithme a été proposé nommé le « Social Page Rank algorithm », qui calcule des scores de popularité permettant de comparer et de discriminer les pages web. (Cai et Li 2010) calcule le degré représentatif des étiquettes tant pour l'utilisateur que pour la ressource pour personnaliser la recherche de l'utilisateur à base d'annotations. Dans le domaine de la recherche bibliographique, un modèle social est proposé pour l'accès aux ressources (Jabeur *et al.* 2010) dans lequel la pertinence d'un document est évaluée selon deux dimensions : thématique et sociale, cette dernière est à son tour dérivée de l'importance sociale des auteurs associés. Deux autres types de relations sont aussi exploités dans cette approche à savoir la relation de citation et l'annotation sociale. Des poids sont attribués aux relations, ils représentent la position des acteurs dans le réseau et évaluent leurs mutuelles collaborations. Dans (Konstas *et al.* 2009), les auteurs combinent les annotations sociales avec les relations sociales entre les utilisateurs pour l'amélioration de systèmes de recommandation. Ces relations sociales sont également exploitées dans (Bouhini *et al.* 2013b) dans but de personnaliser et raffiner la recherche utilisateur. Les auteurs pensent que les utilisateurs peuvent exprimer des besoins d'information différents sous la même requête, ils proposent donc d'intégrer le contexte social de l'utilisateur dans le processus d'indexation des documents pour personnaliser la liste retournée à l'utilisateur.

Plutôt que d'examiner le contenu de la ressource, les auteurs dans (Noll et Meinel 2007; Cai et Li 2010) (Vallet *et al.* 2010) proposent de construire le profil de cette ressource à travers ses annotations et calculent une fonction de correspondance entre le profil de la ressource construit et les termes de la requête, d'une part, et entre ce profil et le profil de l'utilisateur (qui stocke ses annotations personnelles), d'une autre part. Contrairement à cette proposition, (Xu *et al.* 2008) intègrent le contenu de la ressource dans le calcul de fonction de correspondance « ressource-requête » et considèrent que ce contenu doit correspondre aux termes de la requête.

Ces travaux ont montré l'intérêt d'exploiter les informations sociales pour l'annotation des documents, de manière à faciliter la recherche et de la rendre plus efficace. Néanmoins, certains obstacles s'installent avec l'utilisation de ces étiquettes qui sont considérées comme des mots libres et non contrôlés, cela est dû à l'absence de relations sémantiques entre eux, telles que les liens de synonymes, les liens génériques et spécifiques, etc. En conséquence à cala, un problème d'ambiguïté se manifeste lorsqu'une étiquette est liée à plusieurs concepts, on distingue aussi le problème de variations d'écritures lorsque plusieurs étiquettes sont liées au même concept, mais écrites de différentes manières, et le manque d'assistance à l'exploitation de ces folksonomies, notamment pour le stockage, le traitement, et la communication d'information, en l'absence de représentations explicites de ces connaissances (Limpens *et al.* 2008).

Plusieurs approches sont apparues dans la littérature pour atténuer les limitations soulevées dans ce domaine de folksonomies, la plupart d'entre elles balancent entre **i)** la recherche de relations sémantiques entre les termes d'annotation que nous venons de citer (polysémie, synonymie, fautes d'orthographe et variations d'écriture) (Mika 2005) (Specia et Motta 2007) (Angeletou *et al.* 2007) (Buffa *et al.* 2008) (Beldjoudi *et al.* 2011a) (Beldjoudi *et al.* 2012), et **ii)** l'assistance des utilisateurs lors de l'annotation de leurs ressources en leur recommandant les étiquettes qui sont appropriées à leur contenu, celle-ci est connue sous le nom de l'annotation supervisée (Mishne 2006) (Lipczak 2008) (Lu *et al.* 2009) (Pujari et Kanawati 2012) (Lin *et al.* 2011) (Jelassi *et al.* 2014) (Hmimida et Kanawati 2016).

II.2.2.2. Adaptation de la requête utilisateur

Le contenu de la requête utilisateur n'est malheureusement pas toujours optimal pour une recherche pertinente, cela peut être dû à l'utilisateur qui ne sait pas toujours ce qu'il cherche, ou à son manque d'expertise dans le domaine de la recherche, ou tout simplement lorsque cet utilisateur n'exprime pas son besoin de la même façon que l'information interrogée est décrite par le système (cf. section II.2.1). En vue d'améliorer la performance de la recherche qui dépend principalement de cette requête, diverses approches tentent d'améliorer son contenu, certaines d'entre elles essayent de le coordonner avec le langage du système dans l'index, et d'autres aux intérêts personnalisés de l'utilisateur. Pour ce faire, des techniques d'expansion sont adoptées, elles consistent à reformuler ou à enrichir le contenu de cette requête avec de l'information supplémentaire pour donner plus de précision (Zhou *et al.* 2015), et d'autres exploitent ces informations pour affecter le bon sens aux mots ambigus dans cette requête (Navigli 2009), appelées par les techniques de désambigüisation du contenu.

II.2.2.2.1. Expansion de la requête utilisateur

La technique d'expansion est l'une des solutions souvent utilisée pour réduite l'impact de la brièveté et le manque de précision des requêtes des utilisateurs, et au manque d'expertise de ces utilisateurs. Elle permet de générer une nouvelle requête en modifiant son contenu initial par l'ajout de nouveaux termes (Fonseca *et al.* 2005) (Biancalana *et al.* 2008). Il existe une vaste littérature liée à cette technique d'adaptation (Ruthven 2003; Song *et al.* 2007; Wang *et al.* 2009; Lv et Zhai 2010). Ces approches peuvent être classées selon plusieurs critères principaux. Nous citons le degré d'interactivité de l'utilisateur. Ainsi, les approches peuvent être interactives où l'utilisateur représente l'élément principal du processus de reformulation (Ruthven 2003; Fonseca *et al.* 2005), ou automatique lorsque la requête est étendue sans l'intervention directe de l'utilisateur(Gong et Cheang 2006; Song *et al.* 2007) (Carpineto et Romano 2012). Cette extension est faite par l'ajout de termes issus des ressources externes. Selon le principe de génération de ces termes, les approches peuvent être considérées comme linguistiques ou statiques. Les approches linguistiques s'intéressent à la découverte des relations syntaxiques, lexicales, ou sémantiques entre les mots, en s'appuyant sur des ressources terminologiques telles que les ontologies et les thésaurus (Bhogal *et al.* 2007) (Segura *et al.* 2014). Tandis qu'avec les approches statistiques, le principe est de générer des corrélations entre des paires de termes basées par exemple sur la technique de

cooccurrence de termes (Boughareb et Farah 2013a). Cette cooccurrence peut être appliquée sur tout le corpus de documents (Ferber), appelée la cooccurrence globale, ou appliquée uniquement sur les documents retournés derrière la requête de recherche cible, appelée cooccurrence locale. Dans (Kumar et Carterette 2013), les auteurs considèrent le contexte temporel pour évaluer cette cooccurrence en prenant en compte la fréquence des termes qui se trouvent uniquement dans les premiers documents. Le troisième critère sur lequel dépend cette expansion est la source de termes exploitée. Selon cette source de données les approches sont classées en deux catégories, nous distinguons les méthodes basées sur la réinjection de pertinence qui exploitent les feedbacks des utilisateurs évaluant la pertinence des documents, pour construire une collection de données sur laquelle le système se base pour étendre les requêtes des utilisateurs (Salton 1971) (Rocchio 1971). Nous citons aussi les méthodes basées sur les ressources sémantiques. Cette technique se fond sur la sémantique des mots en vue d'identifier leur signification dans un contexte bien précis (Navigli 2009). Les ressources généralement exploitées sont les thesaurus (Park et Ramamohanarao 2007) et les ontologies (Bhogal *et al.* 2007), dont WordNet qui représente la ressource la plus utilisée (Liu *et al.* 2008). Toutefois, pour l'identification des termes adéquats à l'expansion, l'exploitation de telles ressources nécessite le recours aux techniques de désambigüisation de sens des mots.

Le problème majeur avec toutes ces approches est lié à la non-considération du contexte de recherche de l'utilisateur. Pour faire face à cette limitation, d'autres approches se basent sur des techniques personnalisées pour l'expansion de ces requêtes. Elles consistent à enrichir la requête avec les données représentant les intérêts et les préférences des utilisateurs en vue d'adapter le processus de recherche aux besoins spécifiques de chacun. Les données d'intérêt peuvent être de différentes natures, classiques tels que les mots clés extraits des documents d'intérêt de l'utilisateur (Zhou *et al.* 2017), sémantiques telles que les concepts (Audeh *et al.* 2014; Safi *et al.* 2015), ou sociales telles que les étiquettes d'annotation (De Meo *et al.* 2010) (Badache 2013) (Bouhini *et al.* 2016), ou contextuelles telles que la situation géographique ou d'autres propriétés contextuelles (Boughareb et Farah 2013a), etc. Ces données définissent le profil de l'utilisateur, elles sont offertes explicitement par l'utilisateur, ou elles sont extraites implicitement à travers ses interactions avec les systèmes définissant les historiques de

navigation de l'utilisateur (cf. section II.2.2.3.2). La reformulation dans ce cas est guidée par le profil utilisateur et la recherche est considéré comme personnalisée ou dite aussi recherche centrée utilisateur.

Ainsi donc, pour l'expansion personnalisée des requêtes, des auteurs ont examiné les relations entre les étiquettes d'annotation, en sélectionnant celles qui sont plus connexes dans le profil de l'utilisateur (Bender *et al.* 2008) (Bertier *et al.* 2009). Toutefois, les étiquettes ne représentent pas une source fiable pour fournir constamment des descriptions précises des ressources lors de la recherche. Dans les travaux de (Chirita *et al.* 2007) (Biancalana et Micarelli 2009), des analyses locales et des calculs de cooccurrence basés sur le contenu du profil utilisateur ont été adoptées pour étendre la requête. Cependant, le calcul des termes d'expansion est basé uniquement sur une correspondance lexicale entre la requête et les termes dans le profil utilisateur. Les chercheurs dans (Zhou *et al.* 2012b) (Zhou *et al.* 2012a) se basent pour cette expansion sur un modèle qui relie les étiquettes de l'utilisateur à des sujets d'intérêt. Cependant, dans (Zhou *et al.* 2016), les auteurs pensent que la problème majeur avec toutes ces méthodes réside dans le fait qu'elles construisent les profils des utilisateurs uniquement à partir de leurs interactions avec le système. Ils proposent donc une méthode qui permet d'étendre la recherche Web personnalisée en ayant recourt en plus des données sociales de l'utilisateur à une base de connaissances externe pour enrichir son profil.

II.2.2.2.2. Désambigüisation de la requête

Dans la littérature, le problème d'ambigüité peut être linguistique prenant trois formes : forme lexicale polysémique, lexicale homonymique, ou structurale. Ou elle peut être aussi liée au type de besoin de l'utilisateur. En RI, une requête polysémique peut se référer à des documents appartenant à différents contextes d'intérêt. Par exemple pour le dictionnaire BabelNet le mot « sky » a plusieurs significations, il peut être un site des nouvelles de sport, une chaine de télévision, une chaine de radio ou beaucoup d'autres interprétations différentes. Tandis que pour le dictionnaire Oxford ce mot signifie l'atmosphère et l'espace extraits de la terre. La technique de désambigüisation permet d'orienter la recherche vers les documents qui portent sur l'intention de recherche l'utilisateur. Ce processus est une tâche complexe, largement abordée en traitement du langage naturel (TALN). Selon la littérature, les méthodes de désambiguïsation peuvent être divisées en trois catégories (Navigli 2009), basées sur l'apprentissage supervisé, sur l'apprentissage non supervisé, ou des bases de connaissances. Afin de désambiguïser les

nouvelles occurrences des mots polysémiques, la première catégorie utilise des corpus d'entraînement annotés, comportant des étiquettes de sens, en ayant recours à des hypothèses de la théorie de l'information, tels que le modèle Markov caché et Naïve de Bayes. La deuxième catégorie exploite des corpus non annotés pour extraire et inférer les informations nécessaires à la désambiguïsation, et se base sur des méthodes de classification des sens ou clustering. La désambigüisation de sens à base de connaissance quant à elle exploite des ressources sémantiques, elle est basée sur l'idée de calculer la similarité sémantique entre un terme et les autres vocables dans son contexte.

Dans le même ordre d'idée qu'avec la reformulation personnalisée des requêtes des utilisateurs, les techniques de désambigüisation peuvent aussi être guidées par les intérêts spécifiques des utilisateurs sauvegardés dans leurs profils. Cela consiste à apprendre à désambiguïser les requêtes de recherche à partir de leurs intérêts et préférences (Koutrika et Ioannidis 2005) (Jain *et al.* ; Carmel *et al.* 2009; Mihalkova et Mooney 2009; Zhou *et al.* 2012b; Paiva et Ramos-Cabrer 2014).

D'autres catégories contextuelles sont également distinguées dans la littérature pour caractériser une requête de recherche et qui peuvent être utiles pour désambigüiser son contenu. Ces catégories sont liées à la tache de recherche et forment ce qui est connu par le contexte de la tache de recherche. Ce sont les aspects qui sont liés au type du besoin en information de l'utilisateur ou son intention de recherche. Deux aspects contextuels sont distingués : i) l'aspect spécifique qui consiste à associer un ou plusieurs domaines d'intérêt à la requête. Cela a fait l'objet de différentes études, en occurrence, nous citons l'approche proposée dans (Li et Belkin 2008) qui identifie des facettes génériques telles que la durée et le but de la tâche, et des facettes communes qui incluent la perception de l'utilisateur dans la tâche. ii) Et l'aspect général qui consiste à identifier le type de besoin de l'utilisateur selon une taxonomie de trois catégories. Celui-ci peut être informationnel lorsqu'il est lié à la recherche du contenu informationnel de documents, transactionnel lorsqu'il est lié à la recherche des services en ligne, ou navigationnel lorsqu'il est lié à la recherche des sites d'accueil (Broder 2002) (Rose et Levinson 2004; Daoud 2009). La différence entre une requête informationnelle et une requête navigationnelle est que lorsqu'un utilisateur utilise une requête navigationnelle, il attend une seule réponse alors qu'avec la requête informationnelle il attend plusieurs réponses (Lee *et al.* 2005).

Dans le domaine de la RI contextuelle, plusieurs approches ont été proposées pour définir le type de besoin derrière une requête selon cette taxonomie de catégories. Elles visent à mettre en place des techniques de recherche guidées par la tâche de recherche en exploitant les sources d'évidence les plus appropriées à chaque type de besoin (Broder 2002) (Kang et Kim 2003) (Rose et Levinson 2004). D'autres catégories d'ambiguïté sont proposées pour l'identification de l'intention de l'utilisateur derrière sa requête de recherche. Les auteurs dans (Song *et al.* 2009) ont examiné les travaux antérieurs dans (Zhai *et al.* 2003) (Chirita *et al.* 2005) et ils ont construit une taxonomie de trois types de requêtes qui caractérise son niveau d'ambigüité: ambigüe, large, et claire. Une requête ambigüe est une requête ayant plus d'une signification, une requête large est celle qui couvre une variété de sous-thèmes, et un utilisateur peut rechercher l'un des sous-thèmes en émettant une autre requête. Une requête claire est celle ayant un sens spécifique couvrant un sujet étroit. Dans le même cadre d'études, les auteurs dans (Luo *et al.* 2014) proposent un modèle d'identification d'ambiguïté de requête qui prend en compte les caractéristiques comportementales des utilisateurs collectées à partir de leur historique de cliques. Les auteurs se concentrent en plus des fonctionnalités collectées à partir du niveau de requête, sur la façon de distinguer les différences entre des requêtes claires et ambiguës via des fonctionnalités extraites des sessions multi-requêtes. Plusieurs d'autres approches agissent sur les résultats d'une RI pour faire face aux problèmes d'ambiguïté des requêtes.

II.2.2.3. Adaptation de l'accès à l'information

Cette technique d'adaptation consiste à augmenter le processus classique d'appariement avec des données additionnelles pour améliorer et accélérer la recherche. Dans un cadre d'accès personnalisé, les données exploitées sont des informations qui décrivent l'utilisateur pour cibler la recherche à ses intérêts personnels.

II.2.2.3.1. Accès personnalisé à l'information à base du profil utilisateur

L'idée de base de cette personnalisation est d'intégrer la dimension utilisateur U dans la fonction d'appariement qui calcule le score de pertinence des documents interrogés pour une requête de recherche. On écrit : RSV(D,Q) ⇒ RSV(D,Q,U). Selon la littérature, cet appariement s'effectue en deux manières :

- ✓ Exploiter directement le contenu de la dimension utilisateur dans le formalisme de la fonction de pertinence (Lin *et al.* 2005).
- ✓ Exploiter d'une part la structure de la dimension utilisateur et d'une autre part celle des documents interrogés. Ceci, en analysant la structure des relations existantes dans la topologie des documents et en les combinant avec celle du profil utilisateur.

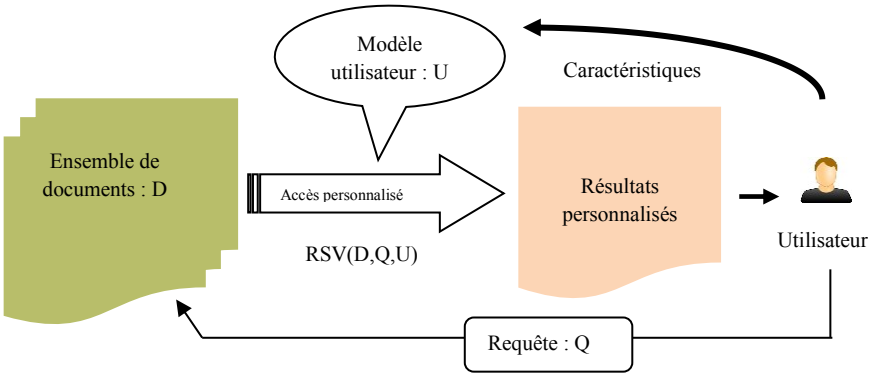


Figure 2. 6. Système de recherche d'information personnalisé (SRIP)

Ainsi, les stratégies de sélection dépendent essentiellement des modèles de représentation des différentes dimensions informationnelles définies dans le processus d'accès, à savoir, le document, la requête, et l'utilisateur. Nous avons présenté dans les sections précédentes différents modèles de représentation et d'enrichissement des documents et de requêtes (cf. section II.1.4). Dans cette section, nous nous focalisons sur le modèle utilisateur qui dépend de deux principales phases, notamment, la collecte de données caractérisant l'utilisateur, et la définition d'une structure de représentation de ces données.

II.2.2.3.2. Collecte des données d'intérêt de l'utilisateur

Cette tâche consiste à extraire les informations qui caractérisent l'utilisateur et qui soient fructueuses pour la construction de son profil et cela dans le but d'améliorer le processus de RI. Ces informations peuvent être de différentes catégories, elles peuvent être i) des données d'identité telles que les données démographiques, le cursus académique, etc. , ii) des données d'activité qui traduisent les interactions de l'utilisateur avec le système, telles que ses historiques de navigation, les évaluations effectuées sur les ressources consultées, etc. iii) des données de préférences telles que la fraicheur d'information, la sécurité, la qualité de service, etc., ou v) des données contextuelles telles que les informations relatives à l'environnement de l'utilisateur, sa localisation, son dispositif, etc. Plusieurs approches utilisent les traces

d'activités, cela consiste à utiliser les évaluations de l'utilisateur effectuées sur l'ensemble des ressources retournées par le SRI (Speretta et Gauch 2005; Tamine *et al.* 2007) (Daoud *et al.* 2009) (Daoud *et al.* 2010b; Mezghani *et al.* 2014). Ces évaluations peuvent être explicites, il s'agit des données rentrées par l'utilisateur (ex. remplissage de formulaire, évaluation par note, évaluation par commentaire, annotation, recommandation, etc.), implicites (ex. impression, sauvegarde, lecture, clics, visites répétées, temps passé sur la page cliquée, etc.), ou externes comme la détection d'un comportement physique dans le cas de la recherche contextuelle dans les mobiles (ex. le mouvement des yeux par exemple, ou autre). Contrairement aux évaluations explicites qui obligent l'intervention de l'utilisateur, celui-ci met en général beaucoup moins d'efforts à noter les choses qu'il aime moins, et trouve généralement que le remplissage manuel de son profil est une tache ennuyeuse, le processus de collecte de données basé sur les deux dernières évaluations (implicites et externes) se distingue avec son mécanisme automatique, en effet aucune information et aucun effort à fournir n'est demandé aux utilisateurs. En revanche, cette technique de collecte de données nécessite une analyse du comportement de l'utilisateur et la pertinence du processus de RI dépend du degré d'efficacité de cette analyse.

II.2.2.3.3. Représentation du profil utilisateur

II.2.2.3.3.1. Modèles de représentation du profil utilisateur

Les données collectées sont exploitées pour la représentation du profil de l'utilisateur sous un modèle utilisateur. Pour ce faire, différentes techniques de représentation ont été utilisées. Selon la structure adoptée, les approches peuvent être classées en 3 catégories: ensemblistes, sémantiques, et multidimensionnelles. Nous commençons par citer les approches ensemblistes qui utilisent une représentation simple et naïve des centres d'intérêt à base de mots clés, tel que le cas des portails web In Quarto, InfoQuest, etc, ou sous la forme d'un ensemble de couples « attribut-valeur » dont chacun représente un sujet d'intérêt et sa valeur. Dans (Cherniack *et al.* 2003), les auteurs utilisent une clause nommée « Domain » qui définit les sujets d'intérêt de l'utilisateur avec un nom abstrait pour chacun, et une clause 'Utility' qui spécifie les valeurs relatives à chaque sujet en utilisant des équations d'utilité. Dans ce cas, un domaine d'intérêt peut avoir plusieurs sujets. De leur part (Bradley *et al.* 2000), le projet est un moteur de recherche d'emploi. Il a pour objectif de personnaliser les offres d'emploi pour chaque

utilisateur selon son profil, ce dernier est représenté sous un ensemble de statistiques qui traduisent les actions de l'utilisateur effectuées sur les annonces du site. Une annonce est considérée comme étant implicitement pertinente pour un utilisateur en fonction du temps de consultation et du type d'action effectuée dessus par l'utilisateur (ex. lecture, candidature, recommandation, etc.).

Des techniques plus élaborées sont proposées, elles se basent pour la représentation du profil sur un ensemble de concepts (Sieg *et al.* 2004) (Daoud *et al.* 2008), ou sur des matrices conceptuelles (Liu *et al.* 2004a), ou selon des vecteurs conceptuels où chacun représente un centre d'intérêt (Mc Gowan 2003) (Tamine-Lechani *et al.* 2008). Selon les travaux de (Daoud *et al.* 2010b) le problème de ces représentations réside dans le fait que ces centres d'intérêt sont déconnectés, aucune relation sémantique n'existant entre eux. Pour combler cette lacune, des approches exploitant des représentations plus complexes ont été proposées, qui relient les centres d'intérêt entre eux. Pour ce faire, dans (Koutrika et Ioannidis 2005), les auteurs utilisent des relations de termes. D'autres regroupent les centres d'intérêt par catégories (Saha *et al.* 2010), ou représentent le profil utilisateur selon une hiérarchie de concepts issue des documents jugés pertinents de l'utilisateur (Begg *et al.* 1993) (Kim et Chan 2003) (Micarelli et Sciarrone 2004). Dans (Sorensen et McElligott 1995), le profil est représenté sous forme d'un graphe orienté où les nœuds sont les mots et les arcs expriment les poids entre ces mots. Chaque poids correspond à une probabilité qu'un mot apparaisse après un autre dans le texte. Dans le même contexte, les auteurs dans (Kießling 2002) représentent le profil sous une hiérarchie de préférences établissant un ordre partiel entre les critères de sélection.

Toutefois, même si ces représentations sont structurées de manière sémantique, le fait que ces centres d'intérêt sont inférés uniquement à partir de l'historique de recherche de l'utilisateur présente une problématique. Celui-ci est souvent limité et ne suffit pas pour détecter un nouveau besoin en informations. Dans le but de remédier à ce problème, des approches de représentation sémantique du profil utilisateur exploitent une ontologie de référence permettant de représenter les centres d'intérêt de l'utilisateur selon un ensemble de concepts pondérés (Liu *et al.* 2004b; Sieg *et al.* 2004) ou une hiérarchie de concepts issue de ladite ontologie (Challam *et al.* 2007; Sieg *et al.* 2007) (Daoud *et al.* 2010b) (Hawalah et Fasli 2015). Il existe plusieurs hiérarchies de concepts et d'ontologies qui permettent de répertorier le web, nous citons la hiérarchie de concepts « MyYahoo » ou celle de l'ODP destinées à

lister et catégoriser les pages web. Ces ressources sont les plus souvent utilisées dans le cadre de représentation de données sémantiques.

Certains d'autres travaux regroupent les informations d'un profil au sein de plusieurs dimensions représentées selon divers formalismes. Dans ce contexte et pour la sécurité, le W3C a défini trois classes pour représenter les profils. Chaque classe possède ses propres attributs liés à un type de données; en occurrence les attributs démographiques, professionnels et comportementaux. Nous trouvons également les travaux (Amato et Straccia 1999a) qui proposent un modèle de représentation du profil structuré en catégories prédéfinies, chacune définit une dimension, en occurrences les données personnelles, de livraison, comportementale et de sécurité. Ce modèle a été conçu dans le cadre du développement d'un service avancé de bibliothèque numérique pour la recherche et personnalisée de l'information. Les auteurs dans (TCHUENTE *et al.* 2012) analysent le comportement de l'utilisateur et décomposent les données collectées en trois dimensions principales: le contenu des données, le contexte des données et la sémantique des données. Dans (Anil *et al.* 2013), les auteurs utilisent les activités en ligne et hors ligne de l'utilisateur pour créer son profil, chacun est stocké dans une dimension, ce modèle considère les variations saisonnières dans l'intérêt des utilisateurs lors de la construction de leurs profils. D'autres auteurs proposent un ensemble riche de dimensions ouvertes dans le sens ou le profil est capable d'accueillir le plus d'informations possible caractérisant l'utilisateur (Kostadinov 2007).

II.2.2.3.3.2. Technique d'enrichissement du profil utilisateur et gestion d'évolution

A. Profil sémantique

Les intérêts des utilisateurs peuvent changer au fil du temps, certains intérêts stockés dans leur profil peuvent devenir obsolètes. Pour faire face à cette limitation, des techniques ont été proposées pour adapter les profils aux besoins courants des utilisateurs. Dans (Zemirli *et al.* 2007; Daoud *et al.* 2010a) (Hawalah et Fasli 2015), les auteurs se basent sur la définition de sessions de recherche par l'introduction de différents paramètres qui permettent la détection du changement dans le contexte de recherche à travers les requêtes utilisateurs, cela permet de définir les intérêts de la session actuelle à travers un profil court-terme utilisé pour le réordonnancement des résultats de la session courante. L'adaptation du profile utilisateur est considérée dans (Zayani *et al.* 2007), par la pris en compte de variables d'intérêt et leur

importance, puis l'introduction de plusieurs caractéristiques évolutives. L'enrichissement des profils utilisateurs peut être aussi fait à travers des techniques d'apprentissage, tels que les réseaux neuronaux, les méthodes de classification (raisonnement par cas, classificateurs bayésiens, etc.), et les règles d'association (Rebaï *et al.* 2013). Ces profils évoluent de plus en plus avec l'évolution des données, et plus particulièrement dans le contexte social, l'utilisateur est de plus en plus actif et représente aussi une source de production de données. À ce sujet, des méthodes exploitent des informations sociales pour enrichir le profil utilisateur comme les métadonnées sur les documents non structurés, à savoir les étiquettes d'annotation (De Meo *et al.* 2010) (Meo *et al.* 2013), la note d'évaluation des documents (Kim *et al.* 2011), les poids des tags (Joly *et al.* 2010), le concept de température qui correspond à la popularité d'un document à un instant donné (Manzat *et al.* 2010) (Mezghani *et al.* 2014). Dans (Abel *et al.* 2011), les auteurs utilisent pour l'enrichissement du profil utilisateur le titre, les auteurs et la date de publication des tweets. Dans le même objectif, d'autres méthodes exploitent le voisinage des utilisateurs (Musiał et Kazienko 2013), cette relation peut être explicite à travers une relation d'amitié, ou à travers des interactions similaires sur le système, déduites en appliquant différentes métriques qui permettent de détecter les communautés d'utilisateurs, comme la similarité cosinus (Kim *et al.* 2011), X-compass (De Meo *et al.* 2010), etc. Pour leur part (Achemoukh et Ahmed-Ouamer 2014), les auteurs s'appuient sur un réseau bayésien dynamique pour enrichir et adapter les profils utilisateurs au fil du temps. Ils ont proposé un cadre théorique qui permet de déduire et d'évoluer le profil utilisateur à partir de ses interactions avec le système de recherche. D'autres chercheurs définissent des catégories d'étiquettes et proposent comment ces catégories peuvent être liées aux dimensions de modélisation de l'utilisateur, incluant l'interactivité de l'utilisateur, les compétences de catégorisation et l'identification de contenu intéressant (Carmagnola *et al.* 2007).

B. Profil social

Le profilage basé sur des étiquettes repose sur l'hypothèse que la pratique de marquage fournisse des informations riches pour créer des profils pertinents pouvant être utilisés pour aider les utilisateurs à trouver des ressources ou des utilisateurs intéressants ou pouvant être utiles pour assister les utilisateurs à utiliser les étiquettes dans les folksonomies.

Les différentes approches exploitant les étiquettes pour la représentation du profil utilisateur peuvent être distinguées par la représentation adoptée pour modéliser le contenu de ce profil, le mécanisme utilisé pour attribuer des valeurs d'importance aux étiquettes, et la stratégie utilisée pour faire face à l'évolution des intérêts. En ce qui concerne le contenu du profil utilisateur les techniques proposées vont d'un simple ensemble d'étiquettes sous forme des vecteurs d'annotations (Yeung *et al.* 2008) (Cai et Li 2010) à des représentations sémantiques plus sophistiquées sous des concepts hiérarchiques (Godoy et Amandi 2008; Hsu 2013), ou un graphique d'étiquettes (Michlmayr et Cayzer 2007).

Ces étiquettes sont extraites en analysant les activités d'étiquetage de l'utilisateur. Dans ce contexte, différents schémas de pondération ont été adoptés tels que la fréquence normalisée de l'étiquette, la fréquence de l'utilisateur inverse, la force de l'étiquette comme le classement, la popularité et la qualité du contenu, etc. L'activité d'étiquetage social ou collaboratif permet d'élargir les profils qui sont généralement limités aux étiquettes personnelles. L'idée derrière cet enrichissement repose sur l'hypothèse que les utilisateurs sont susceptibles de préférer les termes similaires trouvés chez leurs voisinages. Dans ce contexte, Zaho et al., (Zhao *et al.* 2008) supposent que deux utilisateurs sont similaires s'ils partagent plus d'étiquettes fortement liées. Dans (Kim *et al.* 2011) les auteurs présentent une approche qui exploite les étiquettes et leur évaluation en vue de découvrir les sujets d'intérêt pertinents, ces derniers sont enrichis de manière collaborative. De leur part (Michlmayr *et al.* 2007) proposent une approche adaptative étendant un travail antérieur par l'introduction de l'information temporelle qui actualise les poids des arêtes dans le graphe de profil en utilisant une technique d'évaporation. Huang et al. a utilisé deux aspects, à savoir la récente et la durée des balises, pour tenir compte du changement progressif des intérêts des utilisateurs (Huang *et al.* 2014). Dans (Mohamed et Abdelmoty 2016), les auteurs proposent un modèle de géo-folksonomie qui utilise des étiquettes pour des lieux afin de représenter des relations spatiales et sémantiques entre les objets à savoir les utilisateurs, les lieux et les étiquettes. Ce travail examine le problème du profil utilisateur dans les réseaux sociaux basés sur la localisation.

II.2.2.4. Adaptation de l'affichage de données

Cette adaptation se focalise principalement sur l'assistance des utilisateurs à explorer des informations pertinentes. L'idée derrière cette adaptation est de personnaliser l'interface d'affichage des

résultats selon un ou plusieurs critères additionnels en dehors du critère de la pertinence. On distingue les techniques de regroupement thématique des résultats qui permet d'offrir une accessibilité et une navigation plus simple du contenu en regroupant les ressources similaires ensemble qui remplace le mode d'affichage classique en liste. Pour ce faire, différents critères de regroupement sont exploités, à noter les catégories utilisées par les systèmes Vivisimo, Kartoo, Grouper et Exalead, ou les concepts ontologiques dans des domaines spécifiques (Rao et Vatsavayi 2013). On trouve également les approches de classification par taxonomies de concepts adoptées avec les portails web ODP et Google directory qui ont été développées pour faciliter la navigation (Lin *et al.* 1998). Ces techniques se basent sur l'hypothèse que si le contenu d'une ressource est pertinent vis-à-vis une requête, les autres ressources similaires sont peut-être aussi pertinentes.

D'autres approches adaptent l'affichage des résultats aux préférences de l'utilisateur pour personnaliser l'interface, telle que les techniques de réordonnancement et de classement des documents à base du profil utilisateur (Speretta et Gauch 2005; Daoud *et al.* 2010a; Rani 2013), ou à base de son contexte géographique (Geetharani et Soranamageswari 2016), etc.

II.2.3. Systèmes de filtrage d'information

Nous avons vu dans les sections précédentes comment le SRI peut être amélioré selon plusieurs paramètres d'adaptation parmi lesquels nous citons la dimension utilisateur qui permet de filtrer ses résultats selon ses intérêts et ses préférences. Ceci a l'objet de plusieurs formes d'adaptation. Dans cette section, nous discutons une forme spécifique de filtrage d'information orienté utilisateur : la recommandation de données, connue aussi sous le nom de la prédiction d'intérêts utilisateur.

II.2.3.1. Techniques de recommandation de données

Cette section présente les principaux travaux de la littérature sur la recommandation de données, suivis de ceux proposés pour atténuer le problème du démarrage à froid d'un nouvel utilisateur, un problème souvent abordé dans le cadre de la RIP. La recommandation de données est une technique de personnalisation qui permet de guider les utilisateurs durant leur RI en leur proposant les items personnalisés (document web, film, musique, livre, vidéo, image, etc.) les plus susceptibles de les intéresser, ou les étiquettes d'annotation adéquates pour leurs ressources (Jäschke *et al.* 2007) (Musto *et*

al. 2009b) ou pouvant être utiles pour explorer le contenu du système dans un contexte social. Nous distinguons également la recommandation d'individus qui favorise la création de communautés d'utilisateurs ayant des intérêts similaires en recommandant par exemple aux utilisateurs de rejoindre certains groupes (Li *et al.* 2015), ou en leur proposant de contacter d'autres utilisateurs pouvant les intéresser (Jelassi *et al.* 2016) ou des experts sur certains sujets d'intérêt, en occurrence on trouve les systèmes Digg et Flickr qui permettent aux utilisateurs de construire manuellement des groupes sociaux. Pour ce faire, plusieurs techniques sont adoptées, parmi lesquelles on note le filtrage basé sur le contenu dit aussi orienté contenu (Lops *et al.* 2011), le filtrage collaboratif (Su et Khoshgoftaar 2009), et les approches hybrides qui combinent les deux techniques précitées (Thorat *et al.* 2015).

II.2.3.1.1. Filtrage à base de contenu

Le filtrage orienté contenu est une ancienne technique qui consiste à proposer à l'utilisateur cible les ressources qui pourraient lui intéresser en fonction de son profil. Cette technologie est basée sur l'intuition que l'utilisateur introduit des comportements particuliers dans des circonstances données, qui sont susceptibles d'être répétés dans des circonstances similaires. Il s'appuie sur l'évaluation de la similarité entre les ressources (Van Meteren et Van Someren 2000; Domneti 2009) (Lops *et al.* 2011). Dans ce contexte, différents modèles ont été proposés pour représenter ces ressources, nous pouvons citer le modèle d'espace vectoriel à base de mots-clés (Lops *et al.* 2011). Différentes mesures de similarité peuvent aussi être utilisées. La similarité peut être calculée à partir de la mesure de cosinus, ou basée sur la corrélation de Pearson, ou par la similitude cosinus ajustée, etc. En folksonomies, une ressource peut être représentée par l'ensemble de ses étiquettes associées par les utilisateurs et la similarité entre deux items est évaluée par l'appariement de leurs vecteurs représentatifs (Beldjoudi *et al.* 2012). Dans (Durao et Dolog 2012), les auteurs ont enrichi la représentation de documents par divers facteurs basés sur l'utilisation des tags, à savoir leur popularité, leur affinité avec l'utilisateur, et leur représentativité dans le document. Cela permet d'améliorer la sélection des documents pertinents en fonction des données personnelles des utilisateurs représentant leurs données d'utilisation (Gemmell *et al.* 2008) (Durao et Dolog 2012). (Du *et al.* 2016), représentent un document en trois vecteurs: un vecteur des étiquettes favorites, un vecteur des étiquettes non préférées et un vecteur qui englobent toutes les étiquettes. Ces vecteurs sont utilisés pour évaluer la similarité entre les documents. De leur part (Carmel *et al.* 2012) ont

présenté un modèle d'extraction de termes basé sur une folksonomie, qui permet de booster les termes qui sont fréquemment utilisés en public pour annoter le contenu. Dans (Musto *et al.* 2009a), les auteurs ont proposé un service de personnalisation basé sur le contenu afin de trouver les œuvres les plus intéressantes en fonction du profil utilisateur. Ils intègrent les folksonomies en laissant les utilisateurs exprimer leurs préférences pour les ressources avec des évaluations numériques, et des étiquettes d'annotations. Cela a permis au système d'identifier les besoins des utilisateurs et a amélioré les résultats retournés. L'avantage de toutes ces approches réside dans la facilité d'intégration de nouvelles ressources dans le système, ils n'ont pas à être évalués ou marqués par l'utilisateur pour être proposés. Il suffit généralement de leur attribuer la pertinence moyenne estimée de leur voisinage. Aussi, avec cette technique le système ne nécessite pas une grande communauté d'utilisateurs pour être en mesure de faire des recommandations. L'utilisateur peut recevoir une liste de recommandations, même s'il est le seul à utiliser le système. Cependant, cette technique rencontre certaines limitations qui sont énumérées comme suit : 1) elle nécessite l'analyse de chaque nouvelle ressource et cela implique parfois la difficulté d'extraire le contenu ou les attributs de certaines d'elles (données multimédias, documents non structurés, etc.), aussi, la connaissance du domaine est souvent nécessaire, et parfois, des ontologies de domaine sont également nécessaires. 2) Ne peut pas recommander des articles différents lorsque les goûts de l'utilisateur ne sont pas variés, par exemple, si un utilisateur est seulement intéressé par des articles de sport, il ne verra jamais offerts des articles de musique. C'est ce qu'on appelle le problème de sur spécialisation, ce problème ne se limite pas au fait que le système peut ne pas recommander à l'utilisateur des articles différents de ceux qu'il a déjà observés, mais aussi qu'il ne devrait pas lui recommander les articles trop proches de ceux qu'il a appréciés dans le passé, ce qui revient à parler de la diversité des recommandations. 3) pas efficace chez les utilisateurs ayant des goûts variés et non corrélés, 4) pas d'aspect collaboratif dans cette approche.

Afin de faire face aux principaux problèmes du filtrage basé sur le contenu, les techniques proposées ont tendance à utiliser diverses données externes telles que la connaissance approfondie du domaine, les sources de connaissances encyclopédiques, l'exploitation de la richesse des données ouvertes liées (LOD) (Beldjoudi *et al.* ; Di Noia et Ostuni 2015). Ce type de recommandation est bien connu sous le nom de la recommandation basée sur la connaissance et est divisé en deux types: les approches basées sur les

contraintes qui sont fondées sur des règles de recommandation et les approches fondées sur des cas qui sont fondées sur l'identification des ressources ayant des exigences similaires (Jannach *et al.* 2010).

II.2.3.1.2. Filtrage collaboratif

Afin de ne pas limiter l'expérience de recherche pour les utilisateurs individuels, les systèmes de filtrage d'information intègrent des techniques de filtrage collaboratif (FC) pour personnaliser et recommander aux utilisateurs des éléments basés sur leurs communautés. Ces systèmes de filtrage collaboratif (CFS) prennent à l'heure actuelle une place très importante dans le monde des réseaux sociaux et largement investis dans divers domaines. Ils sont conçus pour amener l'utilisateur actuel avec un ensemble d'utilisateurs existants pour lui offrir automatiquement des objets en relation avec ses intérêts aux profils similaires (Beldjoudi *et al.* 2012). Ainsi, si deux utilisateurs Caroline et Sarah ont évalué un certain nombre d'items de façon similaire, il y a de fortes chances que Caroline aime ce que Sarah aime, et inversement. Donc les items que Caroline a aimés peuvent être recommandés à Sarah et inversement. En effet, les CFS se basent sur l'hypothèse que les utilisateurs à la recherche d'information pourraient être intéressés par ce que d'autres ont déjà trouvé et évalué positivement (Su et Khoshgoftaar 2009). Par exemple, dans la vie réelle une personne qui veut lire un livre ou voir un film, demande l'opinion de ses amis qui pourra lui être utile pour faire son choix. Ainsi, un item est d'autant plus pertinent que la proportion d'utilisateurs ayant un profil similaire et ayant apprécié cet item est élevée. Ju et Xu s'appuient sur le regroupement des utilisateurs en utilisant l'algorithme de colonies artificielles d'abeilles pour effectuer une recommandation collaborative (Ju et Xu 2013). Les auteurs dans (Xue *et al.* 2009) ont amélioré la performance de la recherche en développant un modèle de langage d'utilisateur qui utilise les comportements des utilisateurs dans le même groupe pour la recherche collaborative personnalisée. (Cai *et al.* 2014)ont amélioré les méthodes traditionnelles de FC en adoptant l'idée de typicité d'objet dans la science cognitive.

En outre, dans les systèmes d'étiquetage social, l'objectif général de la recommandation de données est d'assurer la quantité et l'adéquation des ressources recommandées. Nous citons le travail de Huang et

al., qui a proposé un système de recommandation qui utilise les étiquettes des utilisateurs les plus récemment identifiés et préférés(Huang *et al.* 2011). (Zanardi et Capra 2011) ont proposé une méthode conçue pour étendre les capacités de recherche des collections numériques visant des domaines universitaires et éducatifs. De leur part, Beldjoudi et al., ont proposé une méthode d'analyse des profils utilisateurs afin d'améliorer la recommandation des ressources (Beldjoudi *et al.* 2011b) (Beldjoudi *et al.* 2012). L'objectif est d'enrichir les profils utilisateurs avec les ressources pertinentes tout en résolvant le problème d'ambiguïté des variables lors de la recommandation.

Cependant, le problème majeur avec tous ces systèmes réside dans le fait qu'ils nécessitent un degré de participation suffisant des utilisateurs en termes d'évaluations de ressources (le cas des systèmes classiques), et d'étiquetage (dans le contexte social), et un nombre suffisant d'utilisateurs. Ce problème est communément connu par le problème dit en anglais « sparcity problem » qui réfère à une situation où les données transactionnelles manquent ou sont insuffisantes. En outre, les nouveaux objets doivent être évalués ou étiquetés avant d'être suggérés, ce problème est connu sous le nom de démarrage à froid d'un nouvel objet. D'autre part, le système n'a pas besoin d'analyser le contenu des éléments à recommander, il n'évalue que la proximité des utilisateurs en fonction de leurs intérêts et propose les éléments associés à ces utilisateurs.

II.2.3.1.3. Technique de recommandation hybride

Ce type de recommandations vient atténuer les limitations rencontrées par les deux techniques abordées dans les deux sections précédentes lorsque chacune d'elle est utilisée individuellement, et ceci par l'introduction de tous les facteurs liés au filtrage des données (utilisateurs, ressources, étiquettes, ressources voisines, utilisateurs voisins, etc.). Nous citons à titre d'exemple le problème de démarrage à froid d'un nouvel objet, rencontré par le FC, ce problème peut être atténué en rapprochant le nouvel objet avec les autres objets du système à travers le FC fondé sur la similarité entre objets. Tel qu'il a été soulevé plus haut, dans certaines situations, le système n'est pas en mesure d'établir la similitude entre les objets non structurés, le FC peut suggérer des objets en se basant sur les évaluations des utilisateurs similaires (Desrosiers et Karypis 2011) . De cette façon, chaque technique peut corriger les limitations spécifiques de l'autre. Actuellement, les SRP les plus efficaces sont basés sur une approche hybride. Plusieurs

approches ont adopté cette hybridation, dans (Lops *et al.* 2013), les auteurs proposent une hybridation basée sur une combinaison linéaire de deux mesures de similarité utilisées. (Gonzalez *et al.* 2007)introduisent des concepts émotionnels spécifiques aux utilisateurs dans un système de recommandation. Le profil utilisateur proposé est composé d'informations issues d'une base de données sociodémographique et des journaux de navigation Web. Lee et ses collègues ont incorporé des données sociales dans le modèle de FC afin de déterminer le nombre d'utilisateurs voisins pouvant être automatiquement connectés sur une plate-forme sociale (Lee et Brusilovsky 2010). Pour améliorer les résultats de la recommandation de ressources au cours de la période de démarrage à froid du système de marquage collaboratif CiteULike. Les auteurs dans (Umbrath et Hennig 2009) ont proposé une approche hybride basée sur l'analyse sémantique latente probabiliste. Par ailleurs, un modèle bayésien à effets mixtes a été proposé par (Condli *et al.* 1999), il intègre les notes des utilisateurs, les caractéristiques des utilisateurs et des ressources dans un seul cadre unifié.

II.2.3.1.4. Réseaux de confiances

Les réseaux de confiances utilisent des techniques semblables au FC et intègrent d'autres critères pour évaluer la similarité entre les utilisateurs. Ils utilisent la relation de confiance entre les utilisateurs qui peut encourager les nouveaux utilisateurs à sélectionner un objet (Jamali et Ester 2009). Cette hypothèse se base sur l'idée que dans la vraie vie, les gens demandent souvent conseil à des personnes de leur entourage à qui ils font confiance, pour choisir un produit commercial, regarder un film, lire un livre, etc. Cette technique a été exploitée dans diverses approches dans le but d'atténuer le problème de démarrage d'un nouvel utilisateur. Ceci représente un principal avantage, car il permet à un nouvel utilisateur de recevoir du contenu en se basant sur son réseau qui se fonde sur les relations de confiance entre utilisateurs. Dans ce contexte (Marsh 1994) distinguent deux types de confiance, une confiance interpersonnelle liée à un contexte précis, et une confiance impersonnelle qui décrit la confiance d'un utilisateur qu'il pour son entourage du groupe. Dans (O'Donovan et Smyth 2005), les auteurs adoptent le FC au sein d'un réseau qui permet à un utilisateur de demander conseil à son entourage lorsqu'il cherche un produit. L'entourage d'un utilisateur est déterminé en évaluant les cotes semblables entre utilisateurs fournis, qui aident à déterminer le degré de fiabilité entre eux. Plusieurs d'autres approches ont adopté l'idée de confiance entre utilisateurs, chacune d'entre elles l'adopte dans un contexte précis. Ces

nombreuses propositions ne peuvent pas être toutes énumérées, pour un aperçu plus détaillé le lecteur est invité à lire les écrits de (Baby et Murali 2016).

II.2.3.1.5. Techniques de détection de motifs

Une autre direction sur laquelle s'orientent les SRIs pour effectuer des recommandations est celle de la fouille des données. Ce domaine inclut un ensemble de techniques visant à extraire des connaissances au sein de grandes collections de données (Han et Kamber, 2006). La détection des motifs est une des techniques de la fouille de données qui peuvent être utilisées pour calculer les recommandations. Pour ce faire, différentes méthodes sont adoptées dont les plus utilisées sont : le regroupement, la classification, la découverte des motifs séquentiels (Adda *et al.* 2007) et les règles d'association. Dans les systèmes de recommandation de données, les règles d'association sont généralement utilisées afin de découvrir les motifs de comportement des utilisateurs qui se répètent dans leurs historiques de transactions avec le SRI.

De manière générale, l'extraction de ces règles nécessite une analyse des actions effectuées par les individus pour déterminer les éléments qui apparaissent ensemble pour la représentation de la dépendance entre eux. Par exemple, si dans un centre d'achat la plupart des clients achètent des lingettes pour bébé et des couches, puis lors d'un futur achat du lait pour bébé, alors le système pourrait créer une règle qui indique que s'il y a des lingettes et des couches dans un même panier, il est fort probable que l'utilisateur achètera du lait. Elle est de la forme $X \rightarrow Y$, où X et Y représentent un ensemble d'objets. L'ensemble d'objets X est nommé la prémisse de la règle, et l'ensemble Y est la conclusion de la règle. Dans le domaine de la RI, cette règle nous renseigne sur la dépendance entre les objets de contenu consommés par les utilisateurs qui peuvent être de différents types (documents web, étiquettes (Beldjoudi *et al.* 2011b), articles commerciaux, films, livres, etc.).

En général, l'extraction des règles d'association passe par trois étapes:

- **La préparation des données** : il s'agit de sélectionner les données de la base de départ et les transformer en base de données transactionnelle. Chaque transaction de données représente un sous-ensemble d'objets, nommé un itemset, et identifié par un identificateur unique (cf. tableau II.1).

- **L'extraction des itemsets fréquents :** cela consiste à déterminer les itemsets qui apparaissent le plus fréquemment dans la base de données transactionnelle, par rapport à un support minimal fixé par le système (cf. définition II.1). C'est l'étape est la plus coûteuse en termes de temps d'exécution puisque le nombre des itemsets fréquents dépend exponentiellement du nombre d'items manipulés. Pour d items, on a (2^d -1) itemsets manipulés (cf. figure 2.7). Ainsi, pour M transactions dans la base de données, la complexité avec cette étape est de O (N*M) où N est le nombre d'itemsets manipulés (N= 2^d).

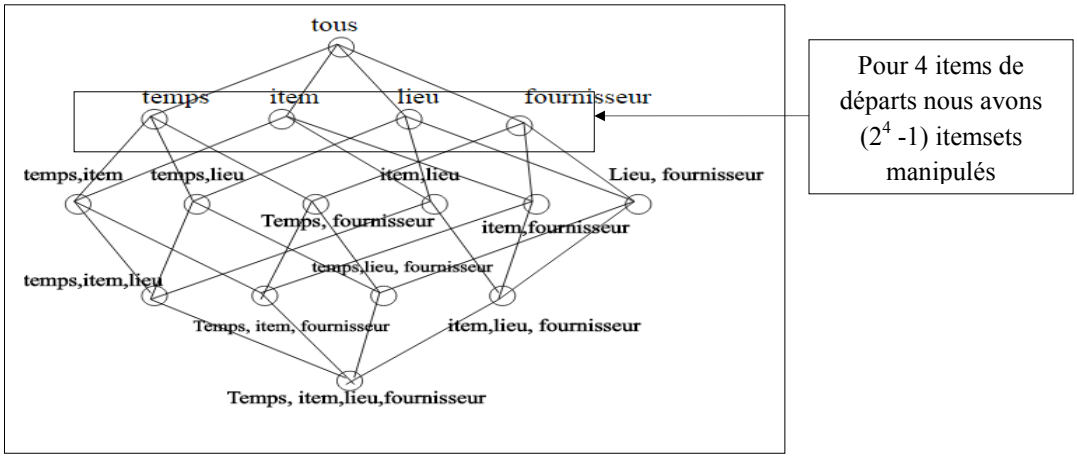


Figure 2. 7. Exemple de treillis d'items

- **Génération des règles d'associations**: À partir de l'ensemble des itemsets fréquents, le système génère les règles d'association qui vérifient un seuil de confiance minimal (cf. définition II.2). Plusieurs générateurs de règles d'association ont été proposés dans la littérature, le premier est l'algorithme Gen-règles, il a été proposé par Agrawal en même temps que l'algorithme Apriori. D'autres algorithmes l'ont succédé, en occurrence AOP2, OPUS, Eclat, etc.

Définition II.1. Le support Supp (Xi) d'un itemset Xi est la proportion d'itemsets dans la base de données transactionnelle qui contient Xi.

Définition II.2. La confiance Conf ($X1 => X2$) d'une règle X1 => X2 mesure la proportion de transactions dans la base de données qui contient X1 et X2 parmi ceux qui contiennent X1.

$$conf(X1 => X2) = \frac{\text{Supp (X1U X2)}}{\text{supp (X1)}} \qquad 2.\ 16$$

Il existe une autre mesure pour mesurer la performance d'une règle d'association nommée le lift. Son évaluation consiste à diviser la confiance de la règle par la valeur espérée de la confiance. Par exemple, pour la règle d'association ($X1 \Rightarrow X2$), le calcul de la valeur espérée de la confiance est le suivant :

$$Conf\ esprérée\ (X1 \Rightarrow X2) = \frac{\text{nombre de transactions contenants X2}}{\text{nombre total de transactions}} \qquad 2.\ 17$$

$$Lift(X1 \Rightarrow X2) = \frac{\text{Conf } (X1 \Rightarrow X2)}{\text{Conf espérée } (X1 \Rightarrow X2)} \qquad 2.\ 18$$

Un lift supérieur à 1 indique une bonne corrélation. Ces paramètres sont illustrés dans l'exemple suivant :

Transaction ID	Itemset
1	lait, pain, beure
2	pain
3	lait, pain, confiture
4	beure
5	Pain, beure

Tableau 2. 2. Exemple de données transactionnelles

À partir de cet exemple, nous pouvons extraire la règle: lait => pain avec une confiance conf (lait => pain) égale à 0.5.

$$\text{conf (Lait => pain)} = \frac{\text{Supp (lait,pain)}}{\text{supp (pain)}} = = \frac{2/5}{4/5} = \frac{1}{2} = 0.5$$

Lift (lait => pain) = $\frac{2/2}{4/5} = \frac{1}{0.8} = 1.25 > 1$, cette règle présente une forte corrélation.

Plusieurs approches ont exploité les règles d'associations, les auteurs dans (Sy *et al.* 2016) les exploitent pour la prédiction des valeurs manquantes dans une base de données. Certaines d'autres propositions les combinent avec les techniques de FC en vue d'améliorer la recommandation de données (Lin 2000) (Bendakir et Aïmeur 2006) (Mican et Tomai 2010) (Beldjoudi *et al.* 2012) (Cakir et Aras 2012) (Alsalama 2013; Wanaskar *et al.* 2013) (Manvitha et Reddy 2014). Toutefois, lorsque l'espace de données contient un grand nombre d'objets, un très grand nombre de règles peut être extrait, qui rend les connaissances obtenues inefficaces (Li et Zhong 2006). Ceci exige des paramètres de contrôle tels que le nombre prédéfini de règles et le niveau de confiance associé aux règles (Lin *et al.* 2002). Dans (Li *et al.* 2004), les auteurs proposent un modèle de contrôle de qualité des règles d'associations permettant de réduire le nombre de règles qui n'intéressent pas les utilisateurs. De leur part Wanaskar et ses collègues

présentent une nouvelle approche basée sur des règles d'association pondérées qui permet d'améliorer les approches proposées dans la littérature (Wanaskar *et al.* 2013). Cela est fait en exploitant des connaissances sémantiques aux résultats pour promouvoir les documents les plus pertinents à une recherche.

II.2.3.1.6. Résolution du démarrage à froid d'un nouvel utilisateur

Dans certaines situations, les systèmes de recommandation n'arrivent pas à proposer des objets de contenu aux nouveaux utilisateurs avec des profils vides. Pour surmonter ce problème, plusieurs approches ont été suggérées. On distingue les approches qui utilisent des réseaux de confiance (Haydar *et al.* 2012; Rohani *et al.* 2014). D'autres ont intégré un modèle d'utilisateur avec des réseaux de confiance et de méfiance pour identifier les utilisateurs dignes de confiance (Chen *et al.* 2013; Guo 2013). Bien que de telles approches soient prometteuses, le fait qu'elles se basent sur les relations existantes entre les utilisateurs présente une limitation lorsqu'un utilisateur est déconnecté du réseau social. D'autres approches ont amélioré les technologies classiques de démarrage à froid en exploitant les données disponibles à froid, telles que l'âge, l'occupation, l'emplacement, etc. pour associer automatiquement les meilleures communautés aux nouveaux utilisateurs (Meng *et al.* 2013) (Zhang *et al.* 2013) (Barjasteh *et al.* 2015). Dans (Safoury et Salah 2013), les auteurs ont évalué l'influence des attributs démographiques sur les évaluations des utilisateurs. Cependant, de telles approches exigent un minimum d'informations sur l'utilisateur, qui ne sont pas toujours disponibles. D'autres sources de données ont été utilisées, telles que les opinions des utilisateurs (Almazro *et al.* 2010) (Wang *et al.* 2011), les étiquettes sociales (Zhang *et al.* 2010) (Preisach *et al.* 2010), des agrégats géographiques (Lanzi *et al.* 2012) (Cuong *et al.* 2012) (Cuong et Long 2013), des arbres de décision (Meng *et al.* 2013), des ontologies (Missaoui *et al.* 2007), etc. En instance, nous distinguons le travail de Sun et ses collègues qui s'appuient sur un arbre de décision intégrant des données démographiques pour associer des utilisateurs existants à un nouvel utilisateur (Sun *et al.* 2011). Dans (Missaoui *et al.* 2007), les auteurs s'appuient sur les concepts et les relations ontologiques à différents niveaux d'abstraction pour développer et enrichir l'ensemble des objets candidats à recommander aux utilisateurs. De leur côté (Zhou *et al.* 2011) ont proposé un schéma d'optimisation itératif qui alterne entre la construction de l'arbre de décision et l'extraction du profil latent afin d'affiner progressivement les profils similaires à l'utilisateur cible. (Zaïer 2010) ont proposé une approche basée

sur une discrimination de voisinage entre utilisateurs, pour chaque utilisateur deux groupes de voisins sont sélectionnés (utilisateurs fortement connectés et faiblement connectés). Cependant, le problème avec cette approche est qu'elle s'appuie sur un profil édité manuellement par l'utilisateur. Dans le même objectif, d'autres auteurs ont introduit le problème des nœuds critiques dans un réseau social en détectant les connecteurs importants dans chaque communauté du système, élus comme responsables à l'assistance des nouveaux utilisateurs (Chekkai *et al.* 2012) (Chekkai *et al.* 2013). Cependant, nous pensons que dans la vie réelle un assistant n'est pas toujours disponible pour diriger les nouveaux utilisateurs. Un nouvel utilisateur doit être guidé dans ses recherches d'information par le système même lorsque les représentants ne sont pas disponibles.

II.2.4. Évaluation des SRIs adaptatifs : systèmes personnalisés et sociaux

Cette section concerne l'évaluation des systèmes qui prennent en considération la dimension de l'utilisateur dans la RI. L'objectif de l'évaluation d'un SRI, quelle que soit sa nature, est de mesurer ses performances vis-à-vis du besoin de l'utilisateur formulé par une requête de recherche. Tel qu'il a été discuté dans la section II.1.5, les cadres d'évaluation des SRIs classiques sont basés sur les approches orientées laboratoire, cette technique se base sur l'utilisation d'une collection de tests où les requêtes sont les seules ressources clés qui traduisent le besoin en information de l'utilisateur. De plus, le jugement de pertinence est purement thématique et totalement indépendant du contexte de recherche de l'utilisateur. Ceci ne permet pas de considérer la dimension de l'utilisateur dans le protocole d'évaluation des systèmes, et engendre par conséquent des limitations pour l'évaluation des systèmes de recherche orientée utilisateur, en particulier les systèmes interactifs et contextuels (Dumais 2009). Ces systèmes ont pour objectif de délivrer de l'information pertinente correspondante à différents paramètres contextuels liés à l'utilisateur, tels que son profil qui englobe ses centres d'intérêt, son environnement, ou autres. Ceci a motivé les chercheurs à réfléchir sur des modèles d'évaluation plus adaptatifs à cette dimension.

Les premières tentatives effectuées dans le cadre de cette recherche ont été proposées dans TREC à travers les tâches interactives et HARD. Ces taches intègrent les caractéristiques spécifiques de l'utilisateur dans le processus de RI, appelées les métadonnées utilisateurs. Cette proposition est effectuée en vue d'améliorer la performance du système pour des requêtes difficiles, en particulier les requêtes

courtes et ambigües. Les métadonnées utilisateur englobent des critères tels que la familiarité, la langue du document, le genre du document, etc. Toutefois, ces critères sont un peu restreints et ne permettent pas d'évaluer un SRI intégrant des aspects contextuels plus larges, tels les centres d'intérêt des de l'utilisateur, le comportement utilisateur mobile, ses informations sociales, etc. Cette limitation a conduit donc à l'émergence des approches d'évaluations fondées sur l'utilisation des contextes de recherche extraits réellement ou par simulation. Comme son nom l'indique, la simulation des contextes de recherche consiste à simuler des utilisateurs et leur interaction avec le système. Un tel cadre est proposé dans (Tamine-Lechani *et al.* 2007) (Tamine-Lechani *et al.* 2008), il représente une extension des cadres d'évaluation TREC via l'enrichissement de leur collection de tests par des profils utilisateurs simulés. Pour ce faire, les auteurs se basent pour la création des contextes sur les interactions hypothétiques fournies par les jugements de pertinence de TREC.

Par ailleurs, l'évaluation par utilisation de contextes réels fait appel à de vrais utilisateurs pour une étude de cas basée sur des contextes de recherche et des interactions réelles de l'utilisateur avec le système. Ces utilisateurs interagissent en deux façons différentes pour la préparation des données de test: i) dans le processus de reformulation des requêtes afin de définir celles qui sont reliées à un même besoin en information définissant une session de recherche. La deuxième façon consiste à utiliser une interface de recherche (l'API Google, Bing ou autre) pour formuler des requêtes selon des besoins spécifiques. Dans ce cas, les documents pertinents sont extraits par une analyse du comportement implicite des utilisateurs en vue d'extraire des fichiers logs, tels que l'analyse des clics, la considération du temps passé sur une page, etc. En occurrence, nous citons le cadre d'évaluation proposé par (Anick, 2003), qui permet d'évaluer un modèle de RI sur des données réelles extraites implicitement à travers le comportement des utilisateurs à partir des fichiers logs.

D'autres protocoles d'évaluation ont été proposés afin d'intégrer la dimension de l'utilisateur. Nous citons les travaux de (Sieg *et al.* 2007) qui proposent de simuler le comportement des utilisateurs en construisant des scénarios de recherche en vue d'évaluer le modèle selon des cas d'étude bien particuliers.

II.3. Conclusion: synthèse et présentation des aspects exploités dans cette thèse

Nous avons présenté au cours de ce chapitre les principaux fondements de la RI classique et ses limitations en présence des requêtes complexes, ambigües et imprécises. Cela est relatif à plusieurs facteurs, en l'occurrence, l'inadéquation des deux langages de représentation requête-document, et aussi à l'utilisateur lui-même, notamment, à son niveau d'expertise et de connaissances sur les domaines de ses recherches. Ces principales limitations ont conduit à l'émergence de différentes techniques d'adaptation du processus RI. Celles-ci se distinguent par les catégories d'information qui sont exploitées dans la représentation des ressources système et de la requête utilisateur en dehors de leur contenu initial. Ces techniques avancées font l'objet de la RI adaptative. Bien que ces techniques aient apporté des solutions pour l'amélioration du processus de recherche, certaines d'entre elles présentent des limitations qui peuvent être énumérées comme suit :

- **Impact du niveau d'expertise de l'utilisateur:** les techniques d'adaptation du besoin informationnel de l'utilisateur exprimé par une requête de recherche s'appuient sur la reformulation du contenu initial de cette requête. Cette reformulation peut-être interactive, c'est-à-dire, le processus a besoin de l'interaction de l'utilisateur pour l'ajout des termes d'expansion. Un degré d'expertise de l'utilisateur sur la recherche cible est nécessaire pour l'atteinte d'un résultat pertinent et amélioré. La performance de ce processus de reformulation est donc liée à ce degré d'expertise.
- **Manque de rétroaction explicite de l'utilisateur:** le processus de reformulation de la requête peut aussi être automatique. Il s'appuie pour l'extraction des termes d'expansion sur les feedbacks des utilisateurs (les jugements de pertinence). Cela demande ainsi l'interaction de l'utilisateur, et dépend fortement de l'aptitude de ces utilisateurs à donner des évaluations correctes, et principalement à leur niveau d'interactivité en temps réel.
- **La non-considération de la dimension utilisateur :** Les ressources linguistiques exploitées pour l'enrichissement du contenu (documents et/ou requêtes) ou pour adapter l'affichage des résultats, ne prennent pas en considération les préférences de l'utilisateur et son contexte de recherche dans ce processus d'adaptation. Cela engendre une certaine limitation liée à la pertinence des résultats. Les systèmes peuvent retourner pour la même requête, les mêmes résultats pour différents utilisateurs. Cependant, ces utilisateurs peuvent avoir différents besoins en information.

Exploitation de l'aspect contextuel dans la RI. Pour améliorer la recherche de l'utilisateur, des techniques plus élaborées ont considéré le contexte de recherche de l'utilisateur dans le processus de RI et l'ont introduit à différents niveaux de ce processus. Nous avons vu au cours de ce chapitre que le contexte peut faire référence à plusieurs paramètres. D'une façon générale, ce sont les facteurs qui interviennent dans le processus de RI pouvant influencer positivement ou négativement sur la pertinence de l'utilisateur et celle du système. Ces facteurs traduisent le contexte de la recherche utilisateur et sont déterminés selon les besoins de cette recherche. Dans la RI contextuelle, cette notion de contexte couvre plusieurs dimensions, parmi les éléments les plus traités dans la littérature nous citons les centres d'intérêt de l'utilisateur connus aussi sous le nom du contexte cognitif, le contexte de la requête de recherche, les préférences de recherche en termes de mode de présentation de résultats et de qualité du contenu offert par le système (ex. fraicheur, crédibilité, etc.). Nous citons aussi le contexte temporel et géographique de recherche, et le contexte d'interactions avec le système connu aussi sous le nom de l'environnement de recherche. Celui-ci peut être cognitif, social ou autre.

Ces paramètres contextuels peuvent être classés en deux catégories : le contexte à court terme et le contexte à long terme. La première catégorie inclut des éléments contextuels qui peuvent changer d'une recherche à une autre, tels que la localisation géographique de l'utilisateur, la nature de la tâche de recherche ou le type de besoin, etc. La deuxième catégorie quant à elle inclut les éléments contextuels qui peuvent persister et évoluer dans le temps, tels que les centres d'intérêt, les préférences de recherche, etc. Cette notion de contexte ne se limite pas à ces paramètres, plusieurs taxonomies ont été proposées dans la littérature pour définir un contexte multidimensionnel (Daoud 2009; Djalila 2014).

Dans cette thèse, nous considérons deux principales dimensions du contexte: i) le contexte du système, et ii) le contexte de l'utilisateur (cf. figure 2.8).

1. Contexte du système : il englobe les caractéristiques qui sont liées au SRI, notamment le niveau représentatif des données (documents, requêtes de recherche, profil utilisateur) et le niveau interactionnel qui définit les stratégies de recherche et de navigation.

a. Niveau représentatif: il définit le modèle de représentation des documents ainsi que le modèle d'interprétation d'une requête de recherche. Ce niveau inclut le contexte du document et le contexte de la requête utilisateur.

- **Contexte du document** : ce sont les diverses catégories d'information qui sont définies pour représenter et enrichir un document.
- **Contexte de la requête :** ce sont les diverses catégories d'information qui sont définies pour interpréter la requête de l'utilisateur et adapter son contenu aux besoins de l'utilisateur.

b. Niveau interactionnel : ce sont les caractéristiques d'interaction qui s'associent au modèle de recherche et de navigation des données sur l'interface utilisateur. Ces caractéristiques définissent de leur côté le type d'environnement offert à l'utilisateur (classique, social ou hybride). Ce niveau définit également le type d'informations qui peuvent être utiles pour faciliter l'exploration des résultats de recherche. Ces informations s'associent au modèle navigationnel offert par le système. Elles peuvent être de différentes catégories : facettes de données, valeurs de facettes, menus, liste de documents, liens d'exploration, etc. Ce niveau définit aussi la technique d'extraction des données d'intérêt de l'utilisateur qui aident à améliorer ses recherches.

2. Contexte de l'utilisateur : il englobe les caractéristiques personnelles de l'utilisateur. Trois niveaux sont définis, à savoir le niveau cognitif, social, et temporel.

a. Niveau cognitif : ce niveau englobe le contexte des taches de recherche de l'utilisateur et ses centres d'intérêt qui sont recueillis durant ses activités de recherche. Ce niveau sert d'une part à offrir à l'utilisateur des résultats de recherche personnalisés, et d'une autre part à la création d'un aspect collaboratif entre les utilisateurs par la formation de groupes d'intérêts similaires. Cet aspect collaboratif aide à améliorer la recherche de l'utilisateur en lui offrant de nouvelles expériences de recherche à base des groupes d'intérêt formés.

b. Niveau social : il permet de définir l'aspect social de l'utilisateur qui peut être collaboratif lorsqu'il appartient à un groupe d'intérêt, ou individuel dans le cas contraire. Il définit aussi le rôle social de l'utilisateur dans le système.

c. Niveau temporel : cette dimension vient définir le contenu des deux niveaux précédents en terme chronologique. Cela consiste à annoter temporellement les données d'intérêt de l'utilisateur. Ceci permet de déterminer leur fraicheur à chaque période de temps. Cette dimension temporelle permet aussi de

définir des groupes d'intérêts évolutifs. Puisque, les intérêts des utilisateurs changent au fil le temps, les groupes d'intérêts changent également à leur tour.

d. Niveau fréquentiel : ce niveau permet de définir le type de besoin de l'utilisateur en deux catégories: temporaire et persistant. Un besoin en information est considéré comme persistant lorsqu'il traduit des centres d'intérêt récurrents qui se répètent à plusieurs activités de recherche. Il est considéré temporaire dans le cas contraire. Cette récurrence aide à définir les données de préférence de l'utilisateur en termes de fréquence.

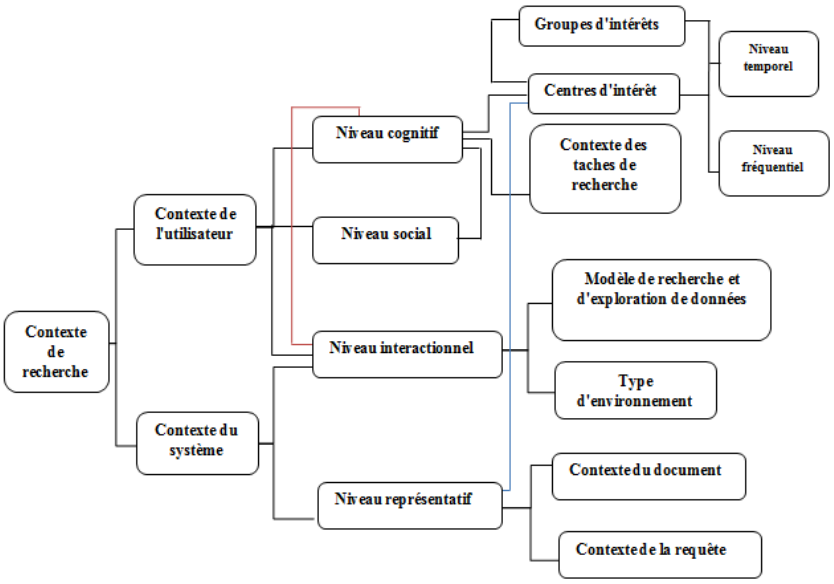


Figure 2. 8. Taxonomie du contexte de recherche proposé

La figure 2.9 ci-dessous illustre les niveaux d'intégration des différentes dimensions contextuelles qui sont définies dans la taxonomie proposée:

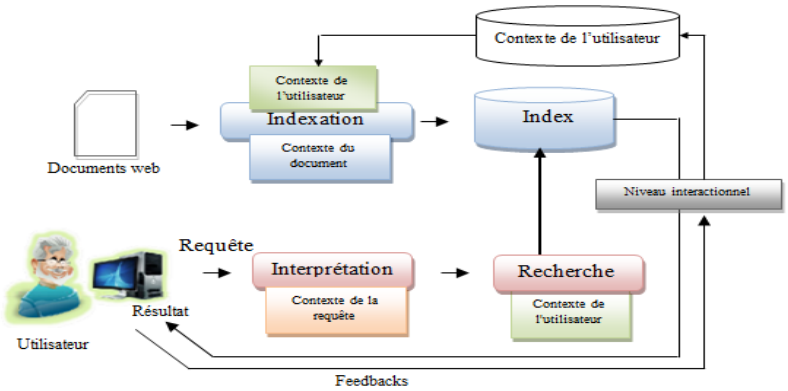


Figure 2. 9. Architecture globale de notre SRI contextuel

Exploitation de l'aspect social dans la RI. Avec l'évolution du web interactif, de nouvelles catégories d'information sont exploitées pour l'enrichissement et l'organisation des données, nous citons les étiquettes d'annotation (folksonomies). Ces étiquettes sont exprimées et introduites librement par les utilisateurs et aident à l'amélioration de la représentation des documents en créant un aspect de représentation évolutif de leur contenu. Ceci contribue à l'amélioration de la RI. Cependant, l'intégration de cette information sociale dans ce processus de représentation a soulevé de son côté différentes limitations liées à l'ambigüité des étiquettes et à leur forme désordonnée**.** Les techniques qui ont été proposées pour remédier à ces faiblesses ont pris deux directions :

- Assister les utilisateurs lors de l'annotation de leurs documents en vue de contrôler le processus d'étiquetage. Ce processus est basé sur l'interaction entre l'utilisateur et le système. Ceci est à l'origine d'une surcharge cognitive pour l'utilisateur, ce qui présente une limite principale pour lui. En outre, l'efficacité du processus d'annotation dépend de la pertinence des termes proposés par le système et aux paramètres qui sont pris en considération pour le choix de ces termes.
- Le rapprochement des folksonomies aux ontologies pour construire des ressources qui aident à relier sémantiquement les étiquettes d'annotation les unes aux autres et aident à désambiguïser leur contenu. L'une des limitations majeures de cette technique est la difficulté de relier les termes entre eux, et l'effort qui se rapporte à cette tâche. Ce problème se produit lorsque les termes sont incompréhensibles par le système. Ceci nécessite l'intervention d'un expert du domaine. Ces limitations restent soulevées et ouvrent la voie à de futures propositions. Elles présentent un des défis de cette thèse.

Exploitation de l'aspect de représentation multidimensionnel. Compte tenu des nombreux avantages qui ont pu être apportés par la recherche multidimensionnelle à base de facettes de données, cet aspect multidimensionnel est intégré dans nos propositions pour la représentation des différentes données qui sont manipulées par le SRI, notamment les documents, les requêtes et le profil de l'utilisateur.

Les prochains chapitres sont consacrés aux contributions de cette thèse. Ces contributions montrent comment les différents aspects précités dans cette section sont exploités au sein de notre SRI.

Chapitre 3 : Nouveau paradigme de recherche d'information sur le Web basé sur un index d'interprétation multi-espaces et un ensemble d'opérations de projection

Ce chapitre présente **i)** un cadre théorique d'un modèle de RI qui offre une représentation réutilisable par les SRI multidimensionnelle (cf. partie 1), **ii)** un modèle d'instanciation qui présente un exemple d'instanciation des concepts clés qui sont abordés dans le cadre théorique (cf. partie 2) et **iii)** une étude de cas qui présente une petite comparaison à petite échelle entre notre système et les moteurs de recherche populaires, notamment Google, Ask et Bing (cf. section III.5).

Partie 1 : Cadre conceptuel d'un modèle de RI multi-espaces

III.1. Introduction

D'une manière générale, l'objectif derrière une modélisation est de permettre une exploitation efficace du processus défini et des données manipulées par ce dernier. Dans notre cas, le modèle représente un processus d'indexation et de recherche multidimensionnelle du contenu web. Le formalisme que nous proposons met en place une représentation qui permet en premier lieu de : i) décrire la structure du contenu web et celle du besoin informationnel de l'utilisateur exprimé par une requête de recherche, et de les enrichir (Adda *et al.* 2013). Puis ii) décrire la manière de chercher dans le contenu web pour trouver l'information pertinente vis-à-vis la requête formulée par l'utilisateur. Cela est fait par la définition d'une opération de mise en correspondance entre les deux univers de représentation, à savoir celui du contenu web et de la requête utilisateur, tout en offrant une flexibilité quant à la manière avec laquelle le contenu web est recherché ou exploré par les utilisateurs. Le but derrière ce formalisme est d'avoir une représentation compréhensible et réutilisable dans divers processus et applications de la recherche et la navigation multidimensionnelle.

III.2. Principaux fondements théoriques

La principale contribution de cette étude constitue à la proposition d'un formalisme rigoureux d'indexation et de RI qui se base sur un nouveau concept appelé la projection multi-espaces (cf. figure 3.1). L'idée principale derrière ce concept consiste à s'appuyer sur des espaces d'interprétation multiples permettant une représentation riche du contenu Web et des requêtes de recherche pour s'adapter aux différents besoins des différents utilisateurs ou d'un même utilisateur. Cette représentation s'appuie sur l'exploitation de descripteurs identitaires du contenu. Ce type de représentation est connu sous le nom de l'interprétation par extraction. Cette représentation s'appuie aussi sur d'autres descripteurs externes extraits de différentes sources de connaissances externes. Cette technique est appelée l'interprétation par association (HUDON 1997). Dans notre système, chaque type de source de connaissance est utilisé pour construire un espace de projection différent. Ainsi, le modèle de description que nous proposons se base sur une hybridation qui combine les deux techniques précitées tout en maintenant les différentes interprétations séparées les unes aux autres en espaces de descriptions. Cette délimitation n'empêche pas l'existence de relations entre les différents espaces d'interprétation qui sont définis.

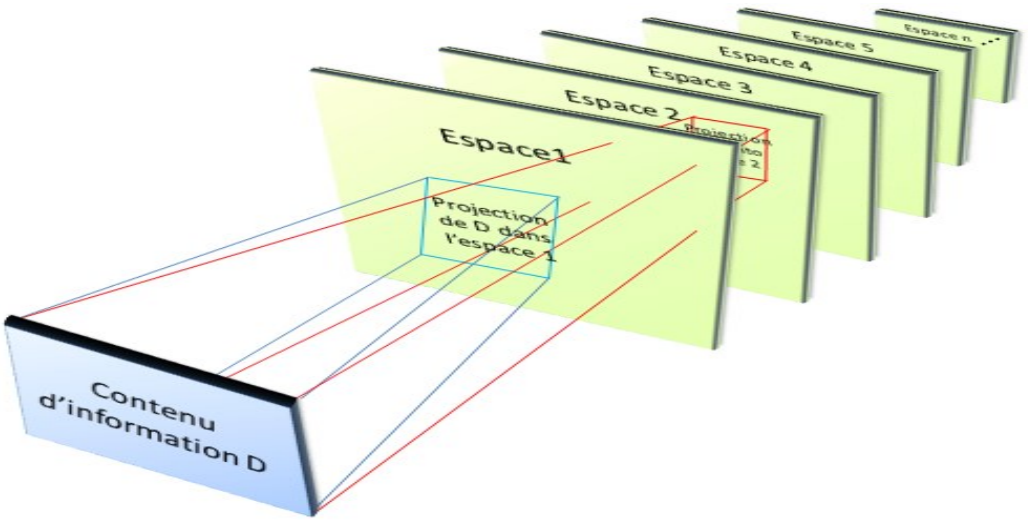


Figure 3. 1. Projection d'un contenu sur les espaces d'interprétation

La projection de données s'effectue par la définition de plusieurs relations à travers l'univers d'interprétation multi-espaces. Chaque relation est basée sur une structure différente et définit une

interprétation spécifique qui peut être instanciée selon le besoin. Chaque espace définit un aspect de description différent qui englobe différentes relations avec le contenu projeté.

Cette projection multi-espaces offre la possibilité de distinguer entre les différents types d'enrichissement appliqués au contenu et ouvre la voie à de nouveaux paradigmes de recherche plus flexibles qui ne sont pas possibles avec les représentations et les enrichissements de données actuelles. Ceux-ci se font généralement sous forme d'un « pèle mêle » de différentes sources et la RI se fait à travers une seule vue de données. Dans notre cas, la RI se fait par l'exploration d'un ou plusieurs espaces d'interprétation qui sont distinguables par les utilisateurs. Chacun contient un sous-ensemble de documents répondant à un type d'interprétation donné.

Le formalisme proposé décrit trois principales tâches en RI, en l'occurrence, l'indexation du contenu web, l'interprétation du besoin informationnel de l'utilisateur, et la mise en correspondance entre les deux univers d'interprétation. Celle-ci est complétée par une technique de navigation. Ces taches font l'objet d'un modèle à deux niveaux: un niveau de définition du système où on s'attache à définir la structure de ses différentes composantes. Ce niveau constitue une vue structurelle du système, il est complété par un niveau qui décrit comment le système réagit suite à une requête de recherche et comment l'utilisateur interagit avec lui pour l'exploration des résultats retournés. Ces différentes taches interactives constituent le niveau comportemental.

III.3. Niveau structurel

Ce niveau définit la structure des différentes composantes du système. Il englobe les concepts clés nécessaires pour l'indexation du contenu web et l'interprétation de la requête utilisateur en vue d'établir le processus de correspondance sensible à cette requête.

III.3.1. Indexation du contenu web

La figure 3.2 présente un aperçu général du processus d'indexation qui se résume comme suit : tout d'abord, le document à indexer est analysé et un ensemble de jetons les plus représentatifs de son contenu (dit en anglais *tokens*) est extrait. Ces jetons sont ensuite projetés sur différents espaces de jetons. Chaque projection est un enrichissement des jetons du document par rapport à un point de vue spécifique

(interprétation). Les jetons résultant de chaque projection sont utilisés pour indexer le document. Les informations sur les espaces qui sont utilisées dans la projection sont conservées dans l'index et exploitées pour chercher et naviguer dans leur contenu.

Ainsi, les principaux éléments manipulés par ce processus d'indexation que nous proposons de formaliser dans cette section sont : le contenu web, les jetons représentatifs du document, les espaces de jetons, et l'index multidimensionnel.

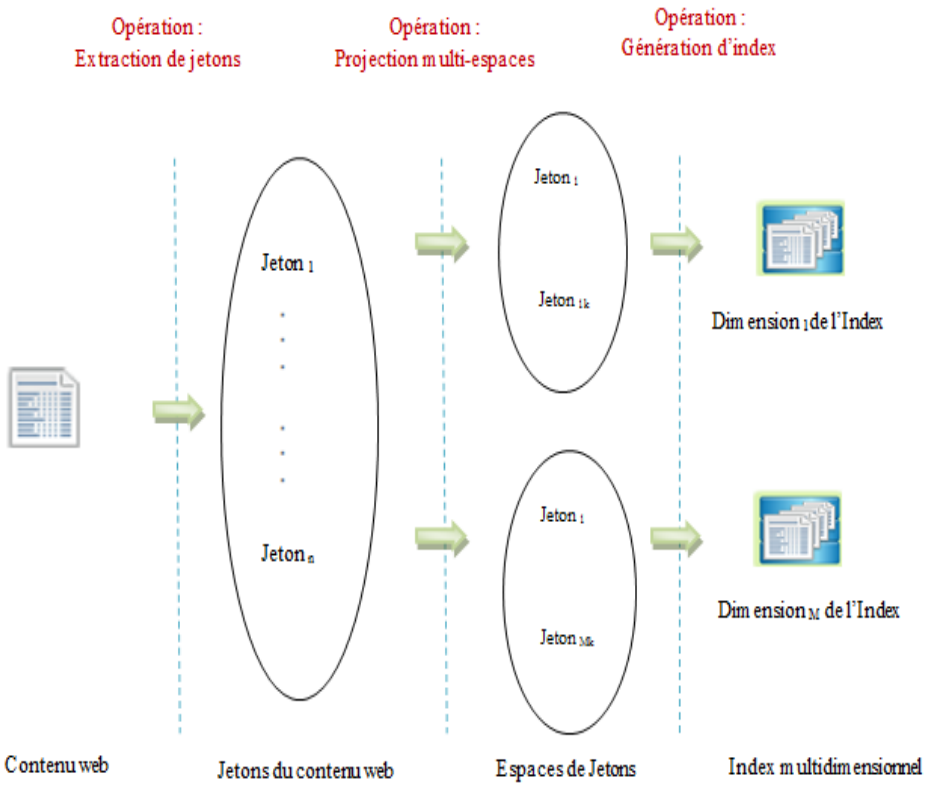


Figure 3. 2. Indexation du contenu web basée sur une projection multidimensionnelle

III.3.1.1. Document web et jetons

Dans notre modèle, le contenu fait référence au contenu des documents web, appelés aussi les pages web. Il est représenté par un ensemble de couples. Chaque couple est composé d'un jeton tk_i et des métadonnées qui le décrivent (cf. définition 3.1) telles que la fréquence d'apparition, la pondération, la position dans le document, la cardinalité, la catégorie grammaticale, la relation syntaxique avec un autre

jeton, etc. Ces métadonnées peuvent être utiles pour extraire le sens local du jeton dans un texte, ou son contexte global dans un ou plusieurs documents. Elles peuvent être également utilisées pour filtrer les jetons qui ne sont pas pertinents, c'est à dire, les jetons qui ne sont pas représentatifs du contenu documentaire (faible pondération) tels que les déterminants, les verbes, ou les jetons ayant une faible similarité avec le sujet du document. Par exemple, un document qui parle sur les virus humains peut contenir des jetons qui n'ont pas de relation directe avec ce sujet tels que «production», «apprentissage», «rapidement», etc.

Définition 3.1. Document Web. Formellement, un document web est représenté par d = {< tk_i, m_i >} tel que ∀ i ∈ [1..n] nous avons :

$tk_i \in U_{str}$ où U_{str} est l'univers des chaines de caractères (str pour String en anglais)

$m_i \in U_m$ où U_m est l'univers des métadonnées

où m_i est une métadonnée qui décrit le jeton tk_i. L'ensemble des jetons tks qui constituent un document d est désigné par $d.tks$. La cardinalité de cet ensemble, notée par $|d|_{tk}$, représente le nombre de jetons qui constituent le document. L'univers de tous les documents web connu sous le nom du corpus documentaire du système est représenté par U_d.

Définition 3.2. **Jeton.** Un jeton tk est la notion élémentaire utilisée dans la représentation du document. C'est un sous-ensemble E de l'univers des chaines de caractères $U_{str}(E \subset U_{str})$. Il est composé d'un ensemble de mots w_j ayant un sens sn_j qui le décrit. Un jeton peut-être de différents types, il peut être une étiquette, un concept, une phrase, une racine de mot, etc. On écrit :

$$tk = \{w_j \text{ pour j } \in [1, |E|]\}$$

Tel que: $w_j \; R \; sn_j$ et R est une relation sémantique qui relit w_i à son sens sn_j

Par exemple, les deux jetons « recherche d'information » et « base de données » représentent chacun un ensemble de mots qu'un système peut utiliser pour décrire un document dans l'index. L'univers de tous les jetons est représenté par U_{tk}.

Un ou plusieurs jetons sont utilisés pour décrire un document d. On écrit [d] pour désigner l'ensemble des jetons décrivant d. Ces jetons peuvent être des éléments de l'ensemble $d.tks$ ou des éléments qui sont en relation de proximité avec à un ou plusieurs éléments de cet ensemble $d.tks$ (cf. définition 3.3). Cette proximité fait référence à plusieurs types de relations qui peuvent relier deux jetons ensemble (cf. définition 3.5 à 3.8).

Pour chaque mot dans le jeton un ensemble de métadonnées inter-jeton est associé. Une métadonnée inter-jeton est similaire à la notion de métadonnées abordée ci-haut. Elle peut être une position du mot dans le jeton, sa relation syntaxique avec un autre mot dans le jeton, sa catégorie grammaticale, un ou plusieurs sens auxquels est lié le mot dans un dictionnaire lexical/sémantique. Cela peut être utile pour déduire le sens approprié du mot au sein du jeton, la décomposition du jeton en sous-ensembles de jetons plus élémentaires, etc. Ainsi, un jeton peut être représenté par un couple tk = {<w_i, m_i>} tel que $\forall$ i ∈ [1,…n] nous avons: $w_i \in U_{str}$ et $m_i \in U_{m.}$.

Définition 3.3. Description du document. Soit un document d∈ U_d et un jeton tk ∈ U_{tk}, on dit que tk décrit d**,** noté par tk ∈[d] si et seulement si ∃ tk' ∈ d.tks : tk' → tk où le signe → se réfère à une relation orientée de proximité entre les deux jetons.

Naturellement, le même jeton peut être utilisé pour décrire plusieurs documents Web. L'ensemble de tous les documents qui sont décrits par le même jeton représente l'ensemble de sa couverture (cf. définition 3.4).

Définition 3.4. Couverture de jeton. Il s'agit de l'ensemble des documents décrit par le même jeton tk. Il est noté par [tk] ([tk] ⊆ U_d). L'ensemble des documents qui représentent la couverture d'un jeton est illustré dans la figure 3.3.

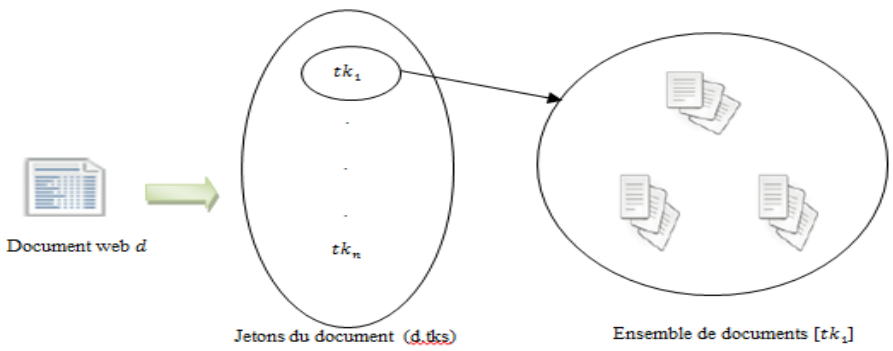


Figure 3. 3. Couverture d'un jeton

Soit un jeton tk = « recherche d'information ». Ce jeton peut être utilisé pour décrire plusieurs documents qui discutent le sujet de la RI. En l'occurrence, nous citons les documents qui couvrent la recherche facettée, ou des livres sur le web sémantique, ou des articles sur la personnalisation des données, etc. Cet ensemble de documents est associé au jeton *tk* et est appelé par « la couverture de la recherche d'information ».

Les jetons peuvent être liés les uns aux autres par différents types de relations d'ordre, parmi lesquels nous citons la relation de généralité (cf. définition 3.5). Puisque ces jetons ne sont pas tous comparables selon cette généralité, on dit qu'ils suivent un ordre partiel en ce qui concerne la relation de généralité.

Définition 3.5. Relation de généralité entre jetons. Un jeton est dit être plus général que l'autre si l'ensemble des documents couverts par ce dernier fait partie de celui du premier. Formellement, cela peut être traduit comme suit : soit deux jetons tk_1 et tk_2 de l'univers U_{tk} $(tk_1, tk_2) \in (U_{tk})^2$:

tk_1 est dit plus général que tk_2 si est seulement si $[tk_2] \subseteq [tk_1]$.

Par exemple, il est évident que les documents Web décrits par « recherche facettée » font partie des documents qui sont décrits par le jeton « recherche d'information ». Ainsi, le deuxième jeton est dit plus général que le premier.

Cette relation de généralité s'appuie sur les ensembles de couvertures des jetons. Toutefois, le calcul de ces ensembles n'est pas toujours pratique, car l'univers du document est, par définition, infini. Ci-après, une relation plus pratique est proposée pour représenter cette relation de généralité. Elle se base sur la structure et la sémantique du jeton plutôt que sur leurs ensembles de couvertures (cf. définitions 3.7 et 3.8). Nous commençons par définir la relation d'ordre dans la définition 3.6.

Définition 3.6. Relation d'ordre partiel entre jetons. Une relation d'ordre r est une relation binaire qui peut ordonner deux jetons appartenant à l'univers de jetons. Étant donné que les paires de jetons n'ont pas besoin d'être toutes liées les uns aux autres, la relation est considérée d'ordre partiel. L'univers de toutes les relations est noté par U_o. Plusieurs instances de r peuvent être identifiées, en l'occurrence nous citons la relation identitaire, de subsomption, de disjonction, etc. Formellement, cette relation d'ordre peut être

définie comme étant une projection $P_{tk}^{S1,S2}$ qui permet d'associer à un jeton *tk* un ensemble de 0 à 1 jeton à travers une relation de différentes catégories (cf. définition 3.12). On écrit : r = $P_{tk}^{S1,S2}$ tel que r ∈ U_o

Définition 3.7. Relation identitaire. Cette relation est un cas particulier de la relation d'ordre entre deux jetons où la correspondance d'un jeton est le jeton lui-même. On dit que tk est une correspondance de lui-même par la relation identitaire r^{ide}. On écrit :

$$\forall\ tk \in U_{str,}\ \exists\ r \in U_o \text{ tel que } r(tk) = r^{ide}(tk) = tk$$

Définition 3.8. Relation de subsomption entre deux jetons. Un jeton tk_1 est dit englobé ou subsumé par un autre jeton tk_2 relativement à une relation d'ordre r ∈ U_o dénoté par $tk_1 \subseteq_r tk_2$ si et seulement si tk_1 précède tk_2 en ce qui concerne l'ordre partiel *o* ($tk_1 \leq_{tk} tk_2$).
Si tk_1 subsume tk_2 alors tk_2 est subsumé par tk_1. Dans ce cas la relation de « est subsumé par » est une relation d'ordre dual à la relation de subsomption.

Il existe d'autres types de relations non ordonnés reliant un ou plusieurs jetons ensemble, tels que la relation de cooccurrence (cf. définition 3.22), la relation contextuelle qui peut être définie selon un paramètre contextuel défini par le système (cf. définition 3.27), etc. L'univers de toutes les relations est noté par U_r.

III.3.1.2. Espace de jetons

Les jetons sont regroupés pour former ce que nous appelons des espaces de jetons ou des espaces d'interprétations, ou simplement des espaces (cf. définition 3.9). Le même jeton peut être trouvé dans différents espaces.

Définition 3.9. Espace de jetons. Un espace de jetons noté par S est un ensemble de jetons tel que S = {tk_i / tk_i ∈ U_{tk} pour tout i ∈ [1..n]}. La cardinalité de S et l'univers de tous les espaces sont respectivement notés par |S| et U_s.

Similairement à l'aspect relationnel qui existe entre-les jetons, des dépendances peuvent être aussi déterminées au sein de l'univers d'interprétation qui sont fondées sur le contenu des espaces. Ainsi, un espace de jetons est dit plus général qu'un autre si son domaine d'interprétation est un sous-ensemble de

celui du second. Puisque l'interprétation d'un espace est l'ensemble des documents qui est couvert par les jetons constituant cet espace, un espace est plus général qu'un autre s'il couvre plus de documents que l'autre dans le sens de la relation mathématique d'inclusion entre ensembles (cf. définition 3.10).

Définition 3.10. Relation de généralité spatiale. Soit S_1 et S_2 deux espaces dans l'univers U_s: $((S_1, S_2) \in U_s^2)$. On dit que S_1 est plus général que S_2 noté par $S_2 \leq_s S_1$ si et seulement la condition suivante est satisfaite :

$$\forall\, tk_1 \in S_1, \exists tk_2 \in tk_2 \text{ tel que } tk_2 \leq_{tk} tk_1$$

Comme c'est le cas avec la relation de généralité entre les jetons, la relation de généralité entre les espaces que nous venons de définir n'est pratique que dans un monde idéal où l'ensemble d'interprétation de chaque jeton dans chaque espace est connu et disponible. Ce qui n'est pas le cas, étant donné le caractère infini des espaces. Ainsi donc, une relation de subsomption spatiale est définie. Elle s'appuie uniquement sur la relation de subsomption entre les jetons (cf. définition 3.11).

Définition 3.11. Subsomption spatiale. Un espace S_2 est subsumé par un espace S_1 (respectivement S_1 subsume S_2), ce qui se note par $S_2 \subseteq S_1$ (respectivement $S_1 \supseteq S_2$) si et seulement si $S_2^I \subseteq S_1^I$ pour toute interprétation I. L'ensemble d'interprétation d'un espace S est défini par l'ensemble des jetons appartement à cet espace. Soit S_1, S_2 deux espaces dans l'univers U_s, $(S_1, S_2) \in (U_s)^2$ on dit :

$$S_2 \subseteq S_1 \text{ si seulement si } \forall tk_2 \in S_2, \exists tk_1 \in S_1 : tk_2 \subseteq_{tk} tk_1$$

Pour illustrer la relation de généralité spatiale, considérons les deux espaces: S_1 = {légume, fruit, véhicule, équipement, compagnie, dispositif} et S_2 = {téléphone, kiwi}. Dans ce cas, l'espace S_2 est subsumé par S_1 étant donné que chaque jeton dans S_1 est subsumé par un jeton dans S_2 relativement à la relation binaire « est- un ». En l'occurrence, le téléphone est un dispositif, et le kiwi est un fruit.

III.3.1.3. Relation de projection et univers de projection

Soit U_s un univers composé d'un ensemble d'espaces, tels que Us = {S_1, S_2, S_3, …, S_n} où S_i de i ∈ [1..n] est un espace de données composé d'un ensemble de points. Chaque point représente un jeton et

peut être projeté à partir d'un espace à un autre espace. D'une manière générale, une relation de projection est la relation qui permet d'associer à un jeton de départ appartenant à l'espace S_1 un ou plusieurs jetons d'arrivée dans l'espace S_2 (cf. figure 3.4) à travers différentes relations d'ordre liant les couples de jetons (cf. définitions 3.5 à 3.8).

Par exemple, le point P_1 = « Java » appartenant à S_1 peut être projeté vers trois points P_1= « langage de programmation » et P_2 = « ile déserte» et P_3 = « magasine» appartenant à S_2. Dans ce cas, la relation de projection qui lie P_1 à ces correspondants P_2, P_3, et P_4 est la relation hiérarchique de subsomption. On dit que java est un langage de programmation, Java est une ile déserte, et Java est un magazine. On écrit: Java$\leq_{\text{est-un}}$ langage de programmation, et Java$\leq_{\text{est-un}}$ ile déserte, et Java $\leq_{\text{est-un}}$ Magazine.

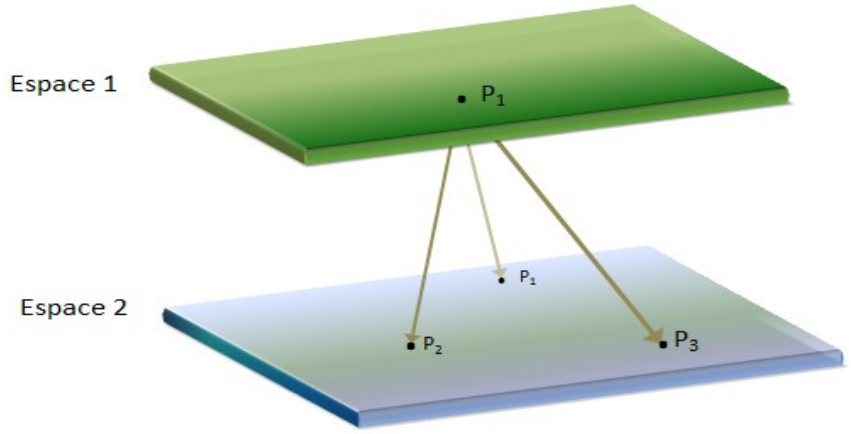


Figure 3. 4. Projection d'un point d'un espace à l'autre

Soit S_1, S_2 deux espaces: $(S_1, S_2) \in (U_s)^2$. Nous commençons par définir formellement la relation projection puis présenter différentes techniques d'exploitation de cette projection.

Définition 3.12. Relation de projection. Formellement, une relation de projection d'un jeton tk, notée $P_{tk}^{S1,S2}$: $S_1 \rightarrow S_2$**,** est une correspondance telle que :

$$\forall \text{ tk} \in S_1, \; P_{tk}^{S1,S2}(\text{tk}) \subseteq \{ tk' / tk' \in (S_2 \cup \emptyset) \; et \; tk' r \; tk : r \in U_r \}$$

où r est la relation qui relie les deux jetons $tk\ et\ tk'$. Elle qui peut être de différentes catégories (cf. définition 3.6 à 3.8).

Un cas particulier de la relation de projection consiste en la projection identitaire basée sur la relation identitaire. Le résultat de la projection est le jeton lui-même.

Définition 3.13. Projection identitaire. Formellement la projection identitaire d'un jeton $P_{tk}^{S1,S2}$: $S_1 \rightarrow S_2$ est une correspondance telle que :

$$\forall\ tk \in S_1,\ P_{tk}^{S1,S2}(tk) = tk$$

L'espace qui nous permet d'obtenir la projection d'identité est appelé l'espace d'identité ou identitaire.

Étant donné qu'un document Web est composé d'un ensemble de jetons, il est considéré comme un espace de jetons spécial. La projection d'un document sur un espace de données est pratiquement similaire à une relation de projection entre deux espaces abordée dans la définition 3.12.

Définition 3.14. Projection documentaire. Soit un document d tel que d = {tk_1, tk_2, tk_3,….. tk_n}, et un espace de jetons $S_1 \in (U_s)$, une relation de projection de d_1 sur S_1 est une correspondance P_d^{S1}: $d_1 \rightarrow S_1$ tel que :

$$\forall\ tk \in P_d^{S1},\ \exists\ tk' \in d_1 :\ P_{tk}^{d,S1}(tk') = tk$$

La représentation d'un document s'effectue par la projection de son contenu sur différents espaces de jetons disponibles (cf. figure 3.1). Cela se traduit par un ensemble de relations de projection entre l'ensemble des jetons qui constitue le contenu du document et les jetons dans les espaces de projection appartenant à l'univers U_s.

Formellement une relation de projection d'un document d_1 sur un univers U_s est une correspondance $P_d^{U_s}$: $d_1 \rightarrow U_s$ telle que :

$$P_d^{U_s} = \bigcup_{i=1}^{|U_s|} P_d^{S_i}(d_1)$$

Définition 3.15. **Projection d'un ensemble de documents**. Soit un ensemble de documents D_1 = {d_1, d_2, d_3, ….,d_n} qui appartient à l'univers des documents U_d ($D_1 \subset U_d$), et S_1 un espace de l'univers U_s($S_1 \subset$

U_s). Une relation de projection des documents notée P_D^{S1} : $D_1 \rightarrow S_1$ est l'union des projections individuelles de chaque document sur l'espace S_1. On écrit :

$$P_D^{S1} = \bigcup_{i=1}^{|D_1|} P_d^{S1}(d_i)$$

Les espaces de projection représentent l'univers descriptif du document qui aide à la génération de l'index documentaire multi-espaces sur lequel se base une recherche multidimensionnelle.

III.3.1.4. Index documentaire multi-espaces

Un index inversé peut être considéré comme étant une structure dynamique où chaque élément est composé d'un jeton, une liste des documents où se trouve ce jeton, appelée la liste d'affectation (concrètement, ce sont les identifiants ids de ces documents qui sont utilisés), et toutes les métadonnées qui décrivent ce jeton, telle que sa position, sa fréquence dans un document donné et sa pondération. Nous commençons par définir la notion d'index simple dit aussi index classique.

Définition 3.16. Index inversé de jeton. Formellement, un index inversé de jetons, noté *I*, est une structure telle que :

$$I = \{tk, List, Meta / tk \in U_{tk} \text{ et } List \subseteq U_d \text{ et } Meta \in U_m\}.$$

Cette structure d'index est flexible dans le sens où elle peut être agrandie par l'intégration de nouvelles informations. La structure telle que définie ci-dessus est proche de la structure des indices inversés habituellement utilisés dans les moteurs de recherche. En effet, en remplaçant les jetons par des mots ou leurs lemmes, on obtient un indice inversé classique. Cependant, cette dernière est rigide, car ne prend pas en charge le modèle de navigation multi-espaces que nous proposons, car les informations sur les espaces de jetons ne sont pas présentes. Nous présentons alors dans ce qui suit une structure d'index qui tient compte des espaces de jetons.

Un index multidimensionnel est un index inversé où chaque entrée de jeton est enrichie avec des informations sur l'espace de son appartenance, telle que la position du jeton *tk* dans l'univers d'indexation U_s noté par «$Pos_{tk}^{U_s}$ » ou le nom de l'espace d'appartenance noté par « NS », etc. (cf. définition 3.17). Comme les caractéristiques binaires (présence ou absence de l'espace de jetons) peuvent être suffisantes

dans certaines situations, elles peuvent aussi ne pas être suffisamment informatives dans d'autres situations où le 0 et le 1 ne suffisent pas pour qualifier l'appartenance d'un jeton à un espace. Ainsi, une valeur est associée à chaque espace de jeton dans l'index, et représente le degré d'association d'un jeton à un espace. Cette valeur peut être interprétée au moment de la recherche selon les besoins de l'application. Par exemple, une application peut être conçue pour renvoyer uniquement des résultats où le poids de jeton est supérieur à un seuil spécifié par l'utilisateur.

Définition 3.17. Index multidimensionnel. Formellement un index inversé multidimensionnel noté par I, est un quadruplet tel que :

$$I = \{tk, List, Meta, <\{< s, i >/ s \in U_{s,} i \in R\}> / tk \in U_{tk} \text{ et } List \subseteq U_d \text{ et } Meta \in U_m\}.$$

Un jeton peut appartenir à des espaces différents, ainsi sa position $Pos_{tk}^{U_s}$ dans l'univers d'espaces U_s est multivalente. Cette position peut être définie comme étant une correspondance qui permet d'associer à un jeton tk_i une ou plusieurs positions dans l'univers U_s. On écrit : $Pos_{tk}^{Us} : tk_i \rightarrow |U_s|$ tel que :

$$\forall\, tk_i \in I\,, Pos_{tk}^{Us}\ (tk_i) \subseteq \{1, .., |U_s|\}$$

Après avoir donné les concepts clés de la représentation documentaire permettant de supporter une recherche multidimensionnelle. Nous passons maintenant à l'élément déclencheur de la recherche, la requête de l'utilisateur. Cette requête représente le besoin en information de l'utilisateur, et son interprétation joue un rôle crucial dans un processus de RI en vue d'offrir des résultats qui répondent au mieux à ce besoin d'information, dits résultats pertinents.

III.3.2. Interprétation de la requête utilisateur

Généralement une requête de recherche est constituée d'un ou plusieurs mots. Dans notre modèle, l'interprétation d'une requête se base sur le concept de jeton défini dans les sections précédentes. Une requête est donc représentée par l'ensemble des jetons les plus représentatifs de son contenu, elle est notée par q tel que $q = \{tk_i / tk_i$ est un jeton pour tout $i \in [1..n]\}$. La longueur de la requête est définie par la

cardinalité des jetons qui la composent, notée par $|q|_{tk}$. L'univers de toutes les requêtes est représenté par U_q .

Une requête de recherche telle qu'elle est formulée par l'utilisateur n'interprète pas toujours précisément son besoin en information, ou peut ne pas correspondre aux jetons qui sont utilisés pour décrire les documents dans l'index. Dans de tels cas, l'adaptation du contenu de cette requête devient primordiale en vue d'améliorer la recherche de l'utilisateur. Pour cette adaptation, différentes techniques d'enrichissement sont souvent adoptées (cf. section II.2.2.2). Elles exploitent différentes catégories d'informations. Ainsi, afin de considérer cette adaptation dans notre modèle, l'interprétation d'une requête utilisateur s'effectue par la projection de son contenu initial sur un univers de description multidimensionnel. Chaque espace correspond à une catégorie différente d'adaptation qui peut être appliquée selon les besoins (cf. figure 3.5).

En outre, un même besoin en information peut être exprimé différemment par différents utilisateurs ou par le même utilisateur. Il peut aussi lié à d'autres besoins en information. Afin de tenir compte de cette réalité, plusieurs relations projectives sont définies à travers l'univers de description de cette requête. Ces relations aident à relier les jetons entre eux et contribuent à l'enrichissement de son contenu pour l'amélioration de la recherche. Cela aide aussi l'utilisateur à explorer le contenu renvoyé par le système suite à cette requête, en fonction de plusieurs interprétations définies à travers les relations projectives.

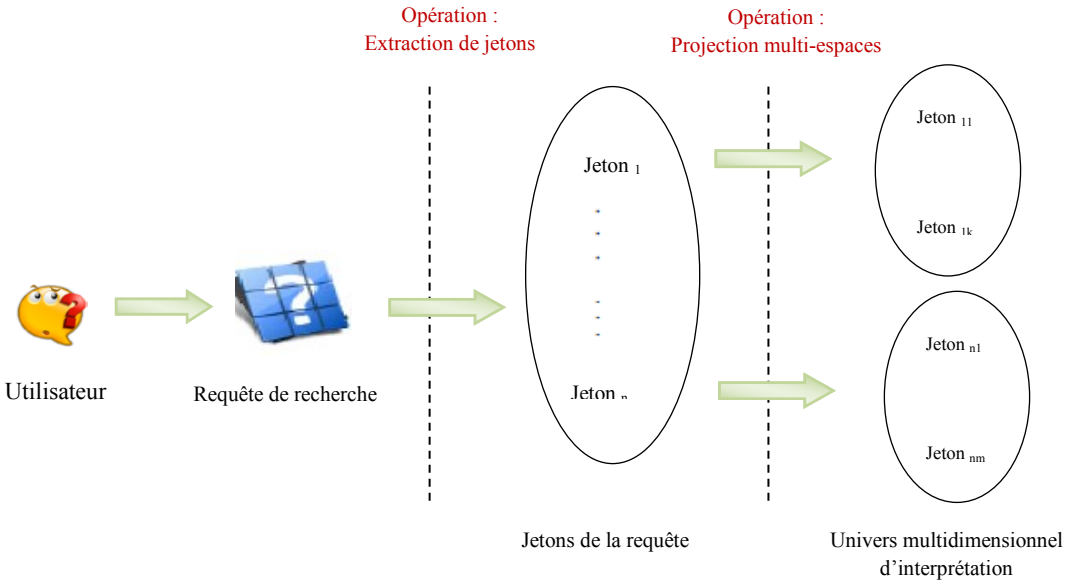


Figure 3. 5. Interprétation d'une requête de recherche

Définition 3.18. Interprétation d'une requête de recherche. Formellement, l'interprétation d'une requête de recherche q consiste en la projection des jetons constituant son contenu initial, lui-même constitué de k-jetons sur un univers de multiples espaces noté par U_s^q. La projection de q vers U_s^q est notée par $P_q^{U_s^q}$: q → U_s^q. Elle est définie comme suit :

$$\forall\, tk_i \in \mathrm{q} : P_q^{U_s^q} = \bigcup_{i=1}^{k} P_{tk}^{U_s^q}(\mathrm{tk_i})$$

L'univers d'interprétation U_s^q inclus la dimension identitaire de la requête représentant son contenu initial, ainsi que les dimensions d'adaptation, qui peuvent être appliquées à ce contenu. Ces dimensions peuvent être de nature contextuelle, sémantique, sociale ou autre. Pour ce faire, plusieurs relations de projection sont identifiées entre les jetons de q et les jetons dans U_s^q, elles traduisent les différentes interprétations possibles de q.

Jusqu'à présent, nous avons présenté les éléments fondamentaux sur lesquels se base un modèle de représentation multi-espaces du contenu. Ci-après, nous présentons le niveau comportemental qui explique comment le système réagit suite à une requête de recherche et comment l'utilisateur peut interagir avec ce système en vue d'explorer pertinemment le contenu proposé.

III.4. Niveau comportemental

III.4.1. Processus de recherche d'information

Un moteur de recherche est destiné à être utilisé pour rechercher des informations en interrogeant l'index documentaire avec la requête de l'utilisateur. Plus concrètement, une recherche associe les jetons de la requête à des entrées dans l'index. Pour chaque jeton interprétant la requête, l'ensemble des documents qu'il couvre est sélectionné pour être retourné à l'utilisateur.

Définition 3.19. Liste de correspondance d'une requête de recherche. Soit q, une requête ($q \subseteq U_q$), E(q) est la fonction qui retourne les jetons de q, A_1 est l'univers d'interprétation qui englobe l'ensemble des espaces de jetons obtenus après la projection multidimensionnelle de E(q), et D_d un sous-ensemble de l'univers des documents U_d ($D_d \subseteq U_d$). On dit que D_d est une correspondance de q si et seulement si

$$D_d = \bigcup_{tk \in A_1}([tk]) \text{ Où } A_1 = \bigcup_{i=1}^{|E(q)|} (P_{tk}^{E(q),U_s^q}(tk_i)) \text{ tel que } tk_i \in E(q)$$

Cette correspondance est basée sur la couverture de jetons (cf. définition 3.4 et figure 3.6). Nous appelons $f(q, I)$ la relation qui permet de déterminer quels documents correspondent à la requête q dans l'index I. Le résultat est un ensemble de documents. On écrit:

$$f(q, I) = D_d \text{ , tel que } D_d = \{id/\ id \in (\{I_q.\ List\} \cup \emptyset) \text{ et } I_q \in E(q)\}$$

Où E(q) est l'univers d'interprétations de q, I_q est une interpretation de q appartenant à E(q) et (I_q. *List*) est la liste de documents qui correspondent à une interprétation I_q dans l'index I.

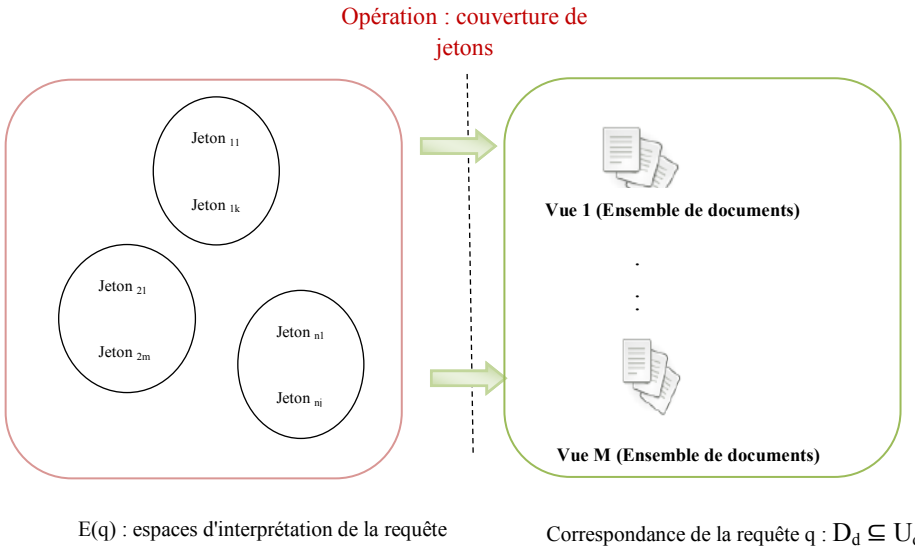


Figure 3. 6. Localisation d'une requête de recherche dans l'index documentaire

III.4.2. Navigation multidimensionnelle

Une technique de navigation définit le système exploratoire offert à l'utilisateur pour exploiter les résultats retournés suite à sa requête de recherche. Dans notre modèle, les informations sur les espaces auxquels appartiennent les résultats de recherche représentent le point essentiel utilisé pour l'organisation de ces résultats sur l'interface de l'utilisateur. Ceci permet de naviguer dans les résultats selon plusieurs interprétations qui sont définies à travers les espaces d'indexation. Cette technique d'exploration est appelée par la navigation multidimensionnelle ou la navigation par facettes.

Définition 3.20. Navigation multidimensionnelle. C'est la façon comment l'utilisateur peut explorer l'ensemble des documents résultant de sa recherche. Cet ensemble de documents provient de différents espaces d'indexation. Chaque sous-ensemble d'un même espace constitue une facette d'exploration fe qui correspond à une interprétation différente de la requête dans l'index (cf. définition 3.21). L'interface utilisateur est dite interface multi-facettes.

Définition 3.21. Une facette de recherche. Une facette de recherche fe est une dimension d'exploration présentée à l'utilisateur pour explorer les résultats de recherche. Elle contient un sous-ensemble de documents $D_{fe} \subseteq D_d$ qui couvrent un ou plusieurs jetons appartenant à l'univers d'interprétation A_1 et correspondant à des entrées dans l'index I ayant la même position « Pos_{tk}^{I} ». Cette position détermine l'appartenance du jeton à un espace d'interprétation S_i et permet de retourner le sous-ensemble de documents qui correspondent à ce jeton au sein d'un même espace d'interopération S_i. Ainsi, une facette de recherche peut être formalisée comme suit :

$$fe = \{ \cup_{i=1}^{|A_1|} [tk_i] \text{ et } tk_i.Pos = k\}$$

Chaque facette peut être enrichie avec des valeurs de facettes qui décrivent son contenu ou traduisent une relation entre son contenu et celui de la requête de recherche. Nous citons la relation de fréquence qui permet d'afficher les jetons les plus fréquents dans un espace de recherche, ou la relation de cooccurrence qui permet d'afficher les jetons qui co-occurrent avec un ou plusieurs jetons de la requête de recherche (cf. définition 3.22) dans le contenu de la facette, celui-ci représenté par l'ensemble de documents D_{fe}. Ainsi, une facette de recherche peut être redéfinie comme suit :

$$fe = \{< D_{fe}, v_i >, v_i \in U_{str} \text{ pour i} \in [1..n]\}$$

Une valeur de facette v_i d'une facette fe est un jeton de description qui peut être exploité par l'utilisateur pour filtrer les résultats de la facette ou explorer le contenu de l'index à travers une nouvelle recherche.

Définition 3.22 : Relation de cooccurrence. Cette relation représente l'apparition simultanée de deux ou plusieurs jetons dans les mêmes documents. On dit que tk_1 est en relation de cooccurrence avec tk_2 si l'intersection de leurs ensembles de couvertures n'est pas vide et la cardinalité de cette intersection dépasse un seuil appelé par le seuil d'apparition et noté Ω. Ce seuil est défini par le système. On écrit :

tk_1 est en cooccurrence avec tk_2 si et seulement si $[tk_1] \cap [tk_2] \neq \emptyset$ et card $([tk_1] \cap [tk_2]) \geq \Omega$.

III.5. Bilan et conclusion

Les principales contributions de ce cadre théorique peuvent être résumées comme suit: nous avons d'abord étendu la représentation classique et monodimensionnelle du contenu web à une représentation multidimensionnelle plus flexible et plus ouverte à de nouvelles stratégies de recherche. Cette nouvelle représentation se base sur la définition de différentes interprétations de ce contenu qui sont maintenues distinguables par le système au sein de l'index et par l'utilisateur au sein de son interface de recherche. Cette distinction est rendue possible grave à la définition de plusieurs espaces d'interprétations et d'un ensemble de relations projectives. Ceci peut répondre aux deux problèmes de manque de diversité sémantique et du contexte de recherche déconnecté, qui ont été soulevés dans le premier chapitre (cf. section I.1.4). Deuxièmement, nous avons montré comment une requête de recherche peut être représentée selon plusieurs dimensions. Ces dimensions peuvent être utiles pour décrire et adapter son contenu selon un ou plusieurs paramètres d'adaptation pouvant être instanciés selon les besoins du système (adaptation sociale, sémantique, contextuelle ou personnalisée). À la fin, nous avons proposé un système de recherche exploratoire basé sur les espaces d'interprétation qui sont définis pour le contenu web et la requête utilisateur. Cette exploration est rendue possible grâce aux facettes de données proposées et aux valeurs de facettes qui peuvent être associées au contenu de chaque facette.

Ce formalisme met en place les bases nécessaires pour développer des SRIs en général et des moteurs de recherche en particulier, qui se basent sur des espaces d'interprétation multidimensionnels. Les fondements théoriques qui découlent de cette proposition peuvent être mis en correspondance avec les fondements orientés objet (OO) comme suit :

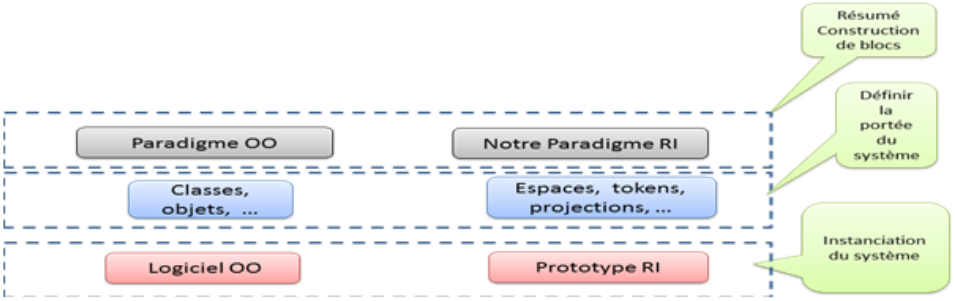


Figure 3. 7. Nos fondements théoriques Vs fondements du paradigme OO

La nouveauté de notre approche est l'utilisation de plusieurs espaces d'interprétations qui sont rendus distinguables les uns des autres de telle manière que l'utilisateur puisse identifier dans quels espaces les résultats sont renvoyés. Ainsi, l'utilisateur peut filtrer ces résultats pour ne conserver que ceux qui sont pertinents à sa recherche. De cette façon, le problème de surcharge de données peut être résolu tout en préservant la sémantique (interprétation) des données indexées et des requêtes des utilisateurs. L'approche multi-espaces que nous proposons est ouverte et flexible dans le sens où les espaces d'interprétation peuvent être définis selon les besoins pour exprimer différentes stratégies d'adaptation. Les opérations de projections peuvent être aussi instanciées et combinées de différentes manières pour exprimer différentes stratégies de recherche.

Pour tester le formalisme proposé, un modèle d'instanciation est nécessaire que nous discutons dans la deuxième partie de ce chapitre.

Partie 2 : Modèle d'instanciation d'un formalisme de RI multi-espaces

III.1. Architecture générale du système

Cette deuxième partie discute le modèle d'instanciation qui emploie les fondements théoriques proposés (Hannech *et al.* 2015). Cela consiste à instancier les concepts clés abordés dans le cadre théorique, notamment les relations de projections (cf. section III.3.1.3), les espaces d'interprétation relatifs au contenu des documents et de la requête de recherche (cf. section III.3.1 et III.3.2), ainsi que la relation de correspondance « requête-document » qui se base sur la couverture de jeton (cf. section III.4). Comme illustré dans la figure 3.8, le modèle d'instanciation proposé comprend les trois principaux modules qu'on retrouve dans les moteurs de recherche traditionnels, en l'occurrence, l'exploration de données, l'indexation de données, et la recherche d'information. D'autres tâches spécifiques à chaque module sont également discutées dans cette partie.

III.2. Exploration et préparation de données

L'explorateur de données, appelé en anglais « the crawler», consiste à naviguer automatiquement sur les sites Web et à télécharger des pages d'intérêt pour alimenter l'indexeur. Il consiste à générer les données qui sont utilisées comme entrée par le module d'indexation. Étant donné que dans notre modèle, les données se réfèrent au contenu des documents Web, la préparation des données est primordiale. Elle consiste à extraire à partir du code html des pages Web explorées, le contenu textuel des documents. Il s'agit d'extraire les métadonnées qui caractérisent ce contenu, tel que le titre, les mots-clés connexes du document s'il existe, et le contenu principal extrait généralement depuis la balise <body> ou <contenu> d'une page html. Il existe également d'autres métadonnées fournissant des informations sur le ou les auteur(s) du document, d'autres liens vers d'autres pages, etc.

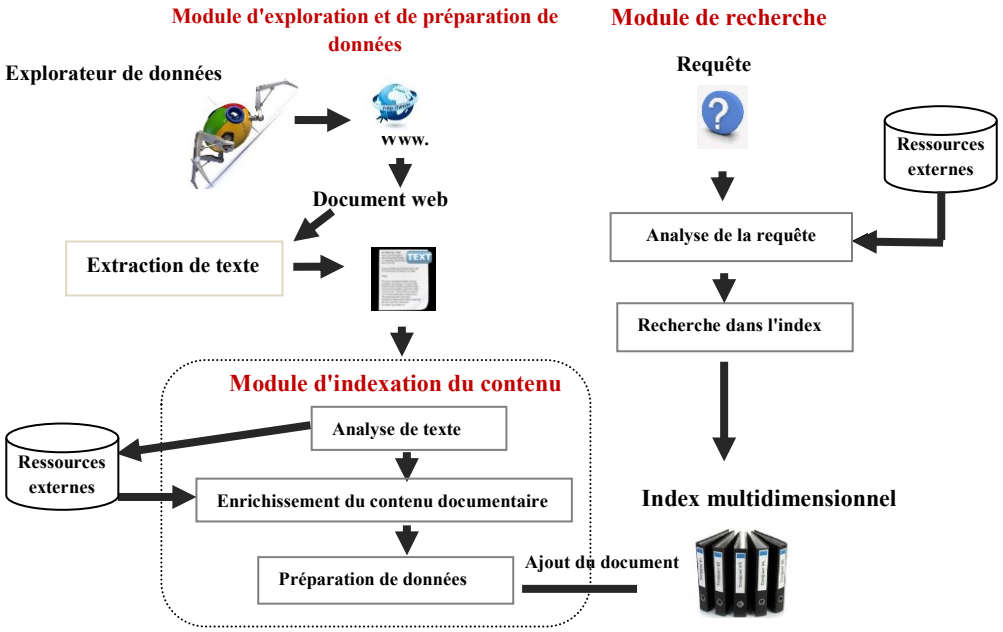


Figure 3. 8. Architecture générale du système

III.3. Indexation des documents web

Pour résumer ce qui a été discuté dans la partie précédente, l'indexation est effectuée en trois principales étapes, i) l'extraction des jetons les plus représentatifs du contenu, ii) la projection de ces jetons sur l'univers de représentation, iii) ces différentes interprétations contribuent à la génération de l'index multidimensionnel qui constitue la troisième étape de l'indexation.

L'instanciation de l'univers descriptif des documents fait l'objet de trois espaces d'interprétation distincts: l'espace d'identité ou identitaire, l'espace sémantique, et l'espace social (cf. figure 3.9). Le premier espace représente le contenu des documents au sens propre du terme sans aucun enrichissement ou adaptation. Il inclut les jetons les plus représentatifs des documents résultant d'une étape d'analyse textuelle. Ces jetons sont appelés les jetons identitaires. Le second espace comprend une liste de jetons qui sont liés sémantiquement aux jetons identitaires, en particulier les synonymes, les hyponymes et les hyperonymes (cf. définition 3.23 à 3.25). Le troisième et dernier espace est l'espace social, il comprend l'ensemble des jetons employés par différents utilisateurs pour annoter les documents à indexer, en l'occurrence les étiquettes d'annotation. Les motifs sur lesquels se fonde cette instanciation sont les suivants :

- ✓ **Pour l'espace identitaire :** une instance originale des documents est gardée, car certains utilisateurs ont une grande préférence pour le contenu qui correspond exactement à leurs besoins exprimés è travers les requêtes de recherche.
- ✓ **Pour l'espace sémantique :** logiquement une information peut être exprimée de plusieurs façons différentes pouvant signifier le même besoin en information (synonymie). Lorsque le contenu de la requête de recherche ne correspond pas au contenu identitaire des documents dans l'index, le recours à d'autres informations qui soient liées sémantiquement à ce contenu peut être utile pour améliorer l'accès aux documents qui peuvent répondre au besoin informationnel de l'utilisateur. Par exemple, un utilisateur est à la recherche des exemples sur des langages de programmation, il tape la requête « langage de programmation ». Le système qui se base uniquement sur l'espace identitaire pour retourner des résultats de recherche ne renverra jamais les documents qui sont indexés avec les jetons « tutoriel Java » ou « tutoriel Perl », en particulier lorsque ces documents ne contiennent pas les jetons de la requête cible, bien que ces documents soient pertinents pour cette requête. Ainsi, afin d'augmenter le rappel du système, il est possible d'enrichir ces documents parlant de Perl et de Java avec le jeton « langage de programmation » qui représente leur hyperonyme (cf. définition 3.25).
- ✓ **Espace social:** l'étiquetage permet à une grande communauté d'utilisateurs d'associer à un contenu plusieurs jetons qui contribuent à son enrichissement. Cette pratique à un avantage double pour notre système. D'un côté, elle permet de créer un aspect évolutif et offrira la possibilité d'ajouter

régulièrement de nouveaux jetons à utiliser pour la description du contenu des documents. Le système peut compter sur une plus grande quantité de vocabulaire à exploiter par l'utilisateur pour la recherche. D'un autre côté, l'exploitation de jetons d'annotation peut aussi être utile pour améliorer la recherche des documents qui ne peuvent parfois être découverts lorsqu'ils ne sont décrits que par des jetons identitaires ou des jetons sémantiques de leur contenu. Par exemple, un document qui discute une nouvelle méthode de RIP et qui soit publié dans une conférence 3W, peut être annoté avec les jetons suivants: nouvelle méthode RIP, article 3W, conférence IEEE. Lorsque ces jetons n'appartiennent pas au contenu identitaire ou sémantique du document, ils peuvent être efficaces pour récupérer ce document.

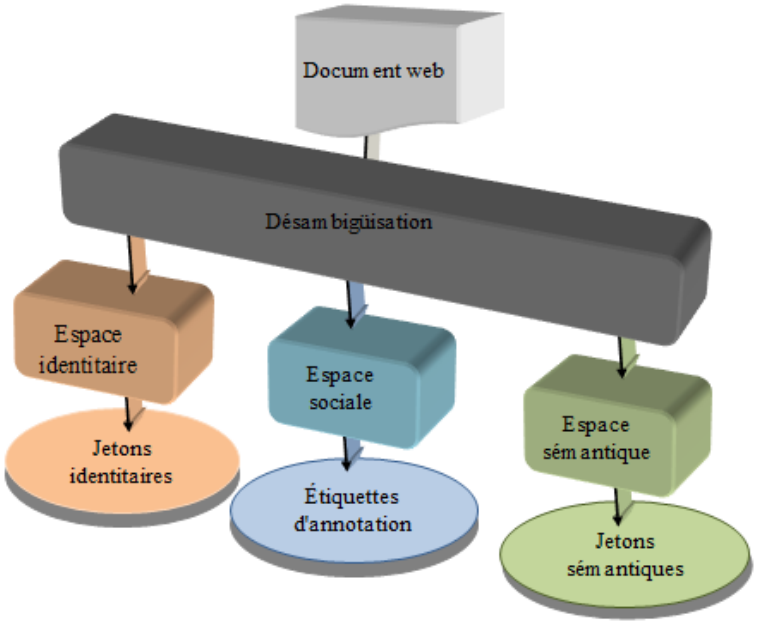


Figure 3. 9. Univers de représentation du contenu documentaire

Définition 3.23. **Synonymie.** C'est une relation qui traduit la similarité entre les jetons en termes de sens/concept. Deux jetons tk_1 and tk_2 sont dits synonymes si et seulement si pour chaque phrase ph_1 qui contient tk_1, alors tk_2 peut être substitué par tk_2 dans ph_1, on écrit : ph_1 $[tk_1/tk_2]$, et pour chaque phrase ph_2 contenant tk_2, alors tk_1 peut être substitué par tk_2 dans ph_2, on écrit : ph_2 $[tk_2/tk_1]$.

Un ensemble de synonymes dénotant le même concept est nommé “Synset”. C’est le composant atomique sur lequel se basent les dictionnaires lexicaux tels que WordNet et BabelNet. Par exemple, les synonymes de « voiture » sont: véhicule, automobile, automobile, etc. Cet ensemble de jetons est un synset.

Définition 3.24. **Hyponymie**. Soit tk_1, tk_2 deux jetons de l’univers des jetons U_{tk}. tk_1 est dit hyponyme de tk_2 si tk_1 est englobé par tk_2 (dénoté par $tk_1 \subseteq tk_2$) si et seulement si tk_1 précède tk_2 en ce qui concerne la relation hiérarchique r «est-un» on écrit : ($tk_1 \leq_r tk_2$). Par exemple, « voiture » est hyponyme de « véhicule ». On écrit : voiture $\leq_{\text{est-un}}$ véhicule.

Définition 3.25. **Hyperonymie.** L’hyperonymie est la relation inverse de l'hyponymie. Ainsi, un jeton tk_1 est dit hyperonyme de tk_2 si tk_2 est hyponyme de tk_1. Par exemple, « fruit » et « nourriture » sont des hyperonymes de « cerises », « couleur » est hyperonyme de « rouge », et « véhicule » est hyperonyme de « voiture », etc.

Il est à noter qu'un jeton peut être lié à plusieurs significations (synsets), on parle dans ce cas de jeton polysémique (cf. définition 3.26).

Définition 3.26. **Polysémie.** La relation de polysémie permet d'associer de multiples synsets à un jeton. Le jeton est dit polysémique. Pa exemple, le jeton « Apple » peut signifier (a) un fruit, or (b) une compagnie, et le jeton « Java » peut être (a) une ile déserte (b) un langage de programmation (c) un logiciel, etc. Cette relation peut être considérée comme étant une correspondance qui permet d’associer un jeton à un ensemble de jetons E_{tk} à travers la relation d'hyponymie « «est-un ». On écrit :

tk est dit polysémique si et seulement si $\exists\ E_{tk}$ tel que $\forall\ tk_i \in E_{tk}\ \ tk \leq_{\text{est-un}} tk_i$ et $\text{card}(E_{tk}) \geqslant 2$

Le processus utilisé pour déterminer le sens d'un mot est connu sous le nom de la désambigüisation. En effet, chaque jeton polysémique doit être désambigüisé en déterminant le sens qui lui correspond selon un contexte local (texte) ou un contexte global (un document ou un sous-ensemble de documents). Cette désambigüisation offre un stockage pertinent du contenu documentaire dans l'index, de sorte que chaque jeton est enrichi avec sa signification, en particulier dans l'espace identitaire. Par exemple, prenons un document Web qui offre des leçons pour apprendre le langage de programmation Java et ayant comme titre le jeton « Java ». Ce document est indexé dans l'espace d'identité avec le jeton « langage Java » au lieu de simplement « Java ». Cela permet d'écarter les documents qui sont indexés uniquement avec le jeton Java et qui ne correspondant pas aux attentes de l'utilisateur lorsqu'il est à la recherche des

documents sur le groupe musical Java ou sur l'ile Java, etc. Cette désambigüisation offre aussi un enrichissement pertinent des documents en associant les bons jetons sémantiques aux jetons identitaires. Par exemple, dans l'exemple précédent le document est enrichi avec les synonymes associés dans un dictionnaire lexical au jeton « langage java » au lieu de ceux qui sont liés aux autres interprétations du jeton « Java » dans ce dictionnaire.

Pour bien illustrer cette démarche, considérons l'exemple suivant. Soit le corpus documentaire suivant :

d_1 = «Java is Indonesia's fifth-largest island. Its 130 million people make up 65% of Indonesia's entire population, and makes Java the most populated island... ».
d_2 = « The Java is a breed of chicken originating in the United States..».
d_3 = « Java is the result of six talented musicians coming together for the single goal of enhancing your event. With lead guitar,..».
d_4 = « Programs written in Java have a reputation for being slower and requiring more memory than those written in C++....».
d_5 = « Welcome to the world of Java examples, organized by categories and Java packages. Java examples (Java sample source code) help to understand .....».

Dans notre modèle, l'index inversé utilisé pour représenter les documents prend la forme illustrée dans la partie A du tableau 3.1, l'index classique prend la forme de la partie B comme suit :

Jeton enrichi	**Doc**
Java dance	D_3
Java island	D_1
Java programming	D_4, D_5
Java chicken	D_2

A. Index inversé enrichi

Jeton	**Doc**
java	D_1, D_2, D_3, D_4, D_5
Island	D_1
programming	D_4, D_5
dance	D_3
chicken	D_2

B. Index inversé simple

Tableau 3. 1. Index inversé enrichi Vs Index inversé simple

Un utilisateur est à la recherche des documents pour apprendre le langage Java, il tape la requête q = « Java tutorial ». Tel qu'il a été vu dans la partie 1 de ce chapitre, le principe général d'un moteur de recherche est de chercher dans l'index les documents qui couvrent un ou plusieurs jetons de la requête de recherche. La requête est composée de trois jetons « Java », « tutorial », «Java programming». Le jeton «java programming » représente une adaptation du contenu de la requête. Cette adaptation est expliquée dans la section suivante (cf. III.4.1.2). Ainsi, une recherche qui répond à la requête q dans

l'index « A » retourne seulement la liste de documents D_4 et D_5, tandis qu'une recherche dans l'index B retourne les documents D_1, D_2, D_3, D_4 et D_5.

III.4. Processus de recherche d'information

Le processus de RI inclut l'interprétation de la requête utilisateur et l'interrogation de l'index documentaire.

III.4.1. Interprétation de la requête de recherche

Tel qu'il a été discuté dans le cadre théorique de la première partie, la requête de recherche exprimant le besoin en information de l'utilisateur passe également par une étape d'interprétation qui permet au système de mieux appréhender ce besoin. Cette interprétation vise à améliorer la recherche de l'utilisateur et consiste en la préparation de plusieurs instances, chacune représente le contenu de la requête sous une interprétation différente représentée à travers une dimension de contenu. En l'occurrence, la dimension identitaire, sémantique et contextuelle.

- ✓ **La dimension identitaire** : Il consiste à extraire les jetons les plus représentatifs de la requête, cette instance de requête constitue le contenu original du besoin informationnel de l'utilisateur sous forme de jetons élémentaires que nous appelons les jetons identitaires.
- ✓ **La dimension sémantique** : Il consiste à enrichir le contenu de la requête avec des jetons qui sont liés sémantiquement aux jetons identitaires. Ces jetons permettent d'élargir ou raffiner la portée de la recherche. Pour cela différentes relations peuvent être utilisées pour augmenter ce contenu, nous citons la synonymie, les relations hiérarchiques spécifiques et générales telles que l'hyponymie et l'hyperonymie. Pour ce faire, il serait important de lever d'abord l'ambigüité sur les jetons polysémiques en vue d'associer des jetons sémantiques pertinents. La requête construite à travers cet enrichissement est appelée par la requête sémantique.

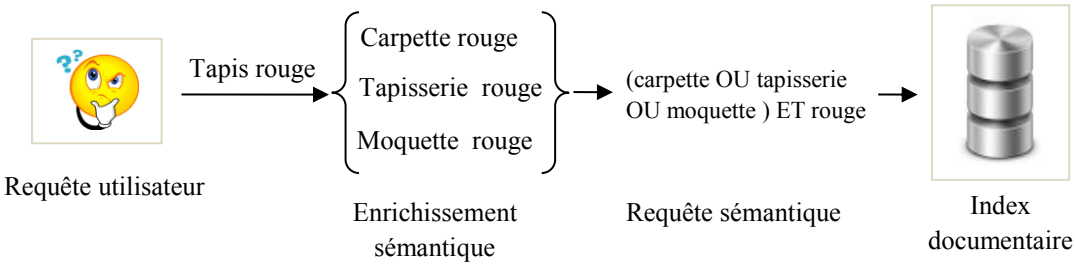


Figure 3. 10. Enrichissement sémantique de la requête utilisateur

✓ **La dimension contextuelle:** Ce sont les requêtes de recherche exprimant le même besoin informationnel de la recherche cible et qui sont formulées différemment. Ce besoin informationnel exprime un sujet de recherche. Ainsi, le système peut utiliser les requêtes soumises par les utilisateurs du système qui portent sur le même sujet de recherche de la requête cible (cf. définition 3.27). Cela permet d'offrir les résultats connexes à une recherche pouvant être intéressants pour l'utilisateur. Pour ce faire, le système exploite l'historique de navigation des utilisateurs pour construire des clusters de requêtes formés à base de sujets de recherche.

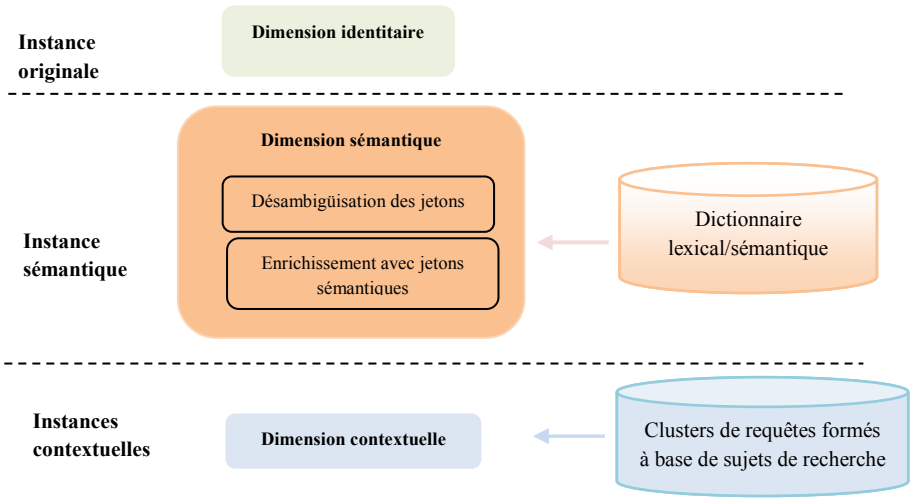


Figure 3. 11. Dimensions d'interprétation de la requête utilisateur

Définition 3.27. Sujet de recherche. Deux activités de recherche sont considérées comme étant liées au même sujet de recherche lorsqu'elles partagent plus de domaines ayant les mêmes rangs d'importance (Daoud *et al.* 2010b).

III.4.1.1. Processus de construction des clusters de requêtes

Le processus de construction des clusters de requêtes est un traitement hors ligne. Il s'effectue au préalable et tourne au fur et à mesure des interactions des utilisateurs, afin d'associer à chaque nouvelle requête un cluster d'appartenance et mettre à jour les requêtes dans l'annuaire des clusters. Ce processus est mis en œuvre en plusieurs étapes : pour chaque nouvelle recherche effectuée sur le SRI, un profil d'activité est construit en projetant sur une ontologie de référence les documents qui ont été jugés explicitement ou implicitement pertinents par l'utilisateur durant cette activité de recherche. Cette projection permet d'attribuer des domaines d'intérêt au contenu de ces documents. Le résultat est un ensemble de domaines d'intérêt pondérés, chaque pondération correspond au score de correspondance entre le contenu des documents avec le contenu d'un domaine dans l'ontologie de référence. Le profil construit est corrélé avec l'historique des activités en vue de l'attribuer à un cluster dans l'annuaire des clusters. Celui-ci présente une forte corrélation avec son contenu (cf. définition 3.27). Chaque cluster de requetes dans l'annuaire est représenté par un seul profil qui englobe tous les profils des activités qui sont fortement corrélées. Cette corrélation est évaluée avec la mesure de Kendall et un seuil de corrélation est appliqué pour retenir les activités de forte corrélation (cf. Annexe 1). Par exemple, les requêtes « langage java », « Java pour les nuls », « java tutoriel », « cours java », et « documentation Java », constituent le même cluster et sont exploitées au sein de la dimension contextuelle.

Nous passons dans la section suivante au processus de désambiguïsation de la requête de recherche lorsque son contenu comporte un jeton polysémique.

III.4.1.2. Modèle de désambiguïsation de sens de mot basé sur le concept de Skyline

Tel qu'il a été déjà relevé précédemment, un ou plusieurs jetons de la requête de recherche peuvent être liés à plusieurs interprétations (sens) qui ne reflètent pas toutes le besoin en information de l'utilisateur derrière cette requête. C'est pourquoi nous proposons une approche multidimensionnelle de désambiguïsation du sens d'un jeton dans une requête. Elle se base sur le principe de similitude de skyline

(Borzsony *et al.* 2001). Ce concept est défini dans notre système comme étant la signification ou le sens le plus similaire à une requête de recherche constituée de plusieurs jetons selon la relation de dominance de Perto.

L'opérateur Pareto a été introduit pour étendre les systèmes de requêtes dans les bases de données (BDD) et rendre efficaces les systèmes de décision multicritères (Borzsony *et al.* 2001) (Yiu et Mamoulis 2007) (Khalefa *et al.* 2008). Il permet de filtrer un ensemble de réponses à partir d'un ensemble de données potentiellement importantes. Pour comparer deux tuples d'une BDD selon une requête de préférence, les valeurs de chaque tuple dans chaque dimension représentant un critère ou un attribut sont comparées par paires. Un ordre partiel peut être défini sur ces tuples en appliquant l'ordre de Pareto (cf. définition 3.28).

Définition 3.28. Relation de dominance. Soit P un ensemble de points multidimensionnels, et X (x_1, x_2, x_3… x_n) et Y (y_1, y_2, y_3...y_n) deux points de P. On dit que X domine Y si et seulement si pour chaque dimension nous avons $Xi \geq Yi$ pour ($1 \leq i \leq n$), et sur au moins une dimension nous avons $Xi > Yi$. On dit que X est préféré à Y et noté $X \succ Y$.

Les requêtes skyline constituent un paradigme très populaire pour extraire des objets à partir d'un ensemble de données multidimensionnelles. Ils sont basés sur le concept de dominance de Pareto. Ils calculent toutes les entrées optimales dans une relation (selon la théorie de l'optimalité de Pareto), c'est-à-dire les entrées qui ne sont dominées par aucune autre entrée dans la même relation (cf. définition 3.29).

Définition 3.29. Le principe de skyline. Le skyline S de P est le sous-ensemble de points non dominés par aucun autre point. Nous notons: $S = \{x_1 \in P / \nexists\, x_2 \in P \text{ tel que } x_2 \succ x_1\}$.

Pour mieux expliquer le concept de skyline, considérons l'exemple suivant: soit une base de données contenant des informations sur les appartements, comme indiqué dans le tableau 3.2 suivant:

Appartement	Surface (M^2)	Années de rénovation
A1	50	2014
A2	35	2009
A3	30	2015

Tableau 3. 2. Exemple de relation d'appartement

Supposons qu'un utilisateur est à la recherche de l'appartement le plus spacieux possible, et ayant récemment été rénové. Nous pouvons vérifier que le Skyline résultant de cette requête est A1 et A3 parce qu'ils ne sont dominés par aucun autre tuple, et A2 est dominé par A1.

Nous tirons profit de cette relation de dominance pour identifier l'interprétation la plus appropriée à un jeton polysémique de la requête de recherche dans une liste de significations possibles extraite depuis un dictionnaire lexical. Les voisins d'un jeton dans une requête fournissent des indices contextuels permettant une désambiguïsation mutuelle de con contenu. L'idée est d'effectuer une comparaison multidimensionnelle en termes de n valeurs de similarité, entre une requête décomposée en n-jetons et une liste de significations (synsets) qui sont liées au jeton polysémique. La similarité est considérée multidimensionnelle. Ainsi, la similarité entre une requête q et un sens M_i est représentée par un vecteur de n valeurs de similarité. On écrit : $SM\ (q, M_i)$ = (v_{i1}, v_{i2}, ..., v_{in}).

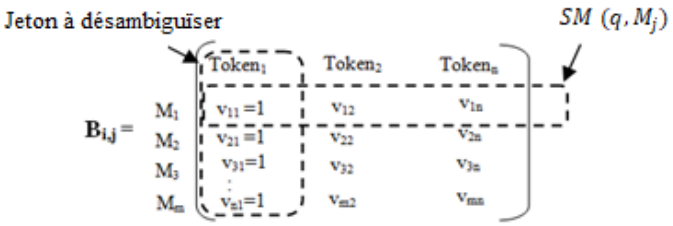


Tableau 3. 3. Matrice de similarité composée

Chaque valeur v_{ij} dans la matrice B_{ij} reflète la similarité entre un synset M_i de la liste des interprétations possibles et un jeton tk_j de la requête. Cette similarité est notée par Sim (M_i, tk_j), elle appartient à un intervalle de valeurs qui varie selon la mesure similarité utilisée. Nous fixons la similarité entre le jeton polysémique et ses synsets à 1 (cf. figure 3.3), car nous considérons que ce jeton est lié à toutes ces interprétations dans un dictionnaire lexical tel que BabelNet et WordNet.

$$\mathbf{B_{i,j}} = \begin{cases} 1, & \text{si } j = 1 \text{ et } i \in [1, n] \\ \text{Sim}(M_i, tk_j), & \text{sinon} \end{cases}$$

Tel qu'il a été discuté dans la section II.2.3.1, plusieurs modèles ont été proposés dans la littérature pour mesurer la similarité ou la distance entre deux mots ou deux concepts. À ce propos, nous avons utilisé pour l'analyse sémantique latente (LSA). Cette mesure a donné dans nos expérimentations de meilleurs résultats comparant à d'autres mesures (cf. section VII.4), en l'occurrence la similarité de Wu et Palmer (Wu et Palmer 1994) et la mesure de Rada (Rada *et al.* 1989).

Une fois que les similarités sont évaluées. Le ou les synsets qui correspondent le mieux au jeton polysémique de la requête sont ceux les plus similaires en termes de dominance Pareto dans la matrice $B_{i,j}$. Ils sont définis par le skyline de $B_{i,j}$. C'est le sous-ensemble des synsets non dominés par aucun autre synset.

$$SYN = \{M_j \in B_{i,j} / \nexists M_k \in B[j] \text{ tel que } M_k \succ M_j\}$$

Le résultat de cette désambiguïsation est le ou les synsets auxquels est lié le jeton polysémique de la requête q. Il est utilisé pour enrichir le contenu de q. Par exemple, la requête q = « Tutoriel Java » est transformée en q' = « Tutoriel Java programming ». Cette nouvelle requête est utilisée pour interroger l'index A dans le tableau 4.1. Le résultat de cette recherche est l'ensemble des documents d_4, d_5 au lieu de la liste d_1, d_2, d_3, d_4, d_5 qui pourrait être retourné suit à la requête originale q soumise par l'utilisateur. Les autres synsets éliminés peuvent être également utilisés pour écarter les documents non pertinents.

III.4.2. Recherche d'information et exploration des résultats

Les différentes instances obtenues de la requête de recherche représentent chacune son contenu selon une interprétation différente. Elles sont exploitées pour interroger l'univers d'indexation des documents. Chacune est utilisée pour retourner un sous-ensemble de documents dans un espace d'affichage distinct (la facette de données). La requête originale est utilisée pour interroger l'espace d'indexation identitaire et retourner dans la facette identitaire les documents qui correspondent à son contenu. Il s'agit des documents dont le contenu identitaire correspond syntaxiquement à un ou plusieurs jetons de cette dimension identitaire de la requête (cf. figure 3. 12 partie 1). Cette dimension identitaire est également exploitée pour retourner à la fois dans la facette sociale, les documents qui ont été annotés par les utilisateurs avec un ou plusieurs jetons de son contenu (cf. figure 3.12 partie 2) et dans la facette sémantique les documents qui sont décrits sémantiquement avec un ou plusieurs jetons de son contenu (cf. figure 3.12- partie 3). La requête sémantique quant à elle est utilisée pour retourner dans la facette sémantique, les documents dont le contenu identitaire et social correspond avec un ou plusieurs jetons de cette dimension sémantique (cf. figure 3.12- partie 3). Ainsi, la facette sémantique contient à la fois des documents qui correspondent au contenu sémantique de la requête dans l'espace identitaire et social, et ceux qui correspondent à la requête originale dans l'espace d'indexation sémantique. Enfin, la dimension

contextuelle de la requête permet de proposer des recommandations liées au même sujet de la recherche cible, ces recommandations sont proposées sous forme de requêtes au sein de la facette sémantique (cf. figure 3.12- partie 3).

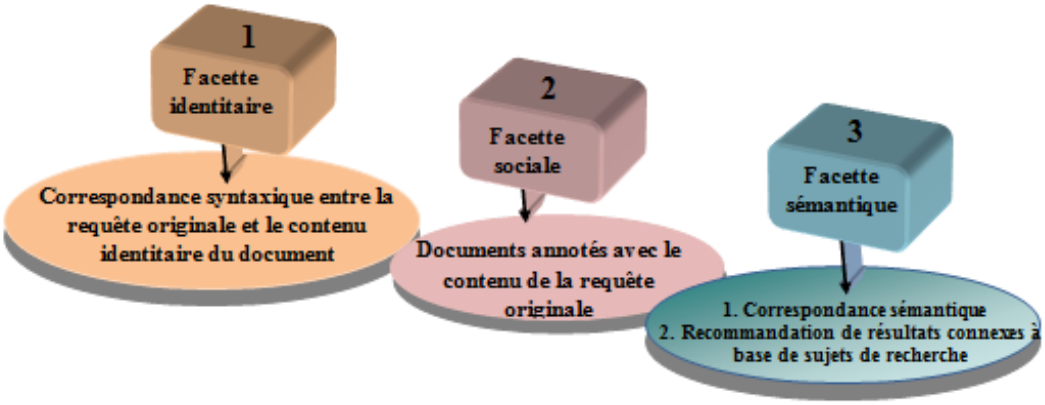


Figure 3. 12. Interface utilisateur multi-facettes

III.5. Études de cas comparatives

Pour évaluer le système proposé, un prototype fonctionnel est développé. Il met en œuvre le modèle d'instanciation présenté dans la deuxième partie de ce chapitre et est maintenu par des données réelles. Ce prototype est présenté dans le chapitre 7 de cette thèse (cf. section VII.2). Une série d'expérimentations est effectuée pour tester et évaluer son efficacité (cf. chapitre 7).

Dans cette section, une petite comparaison à petite échelle est présentée entre notre système et les moteurs de recherche populaires, notamment Google, Ask et Bing. Elle est effectuée en fonction de l'interprétation du contenu et d'utilisabilité du système (fonctionnalités de recherche, facilité d'utilisation et pertinence des résultats offerts).

Prenons le premier exemple d'un utilisateur qui fait une recherche sur « Java ». Ce jeton est lié dans les dictionnaires linguistiques tels que WordNet et BabelNet à plusieurs interprétations (cf. figure 3.16), et l'utilisateur ne s'intéresse probablement qu'à une seule d'entre elles. Si nous interrogeons un moteur de recherche comme Google, Bing ou Ask.com, les seules interprétations proposées à l'utilisateur dans les premières pages sont liées à la plate-forme Java ou au langage de programmation Java, et cela continue à la septième page et au-delà (cf. figure 3.13 partie 3). C'est comme si ces moteurs de recherche font l'hypothèse que les utilisateurs sont généralement intéressés par le langage de programmation Java et

l'écosystème (cf. figure 3.13 parties 1 et 2). Cette hypothèse peut être fondée sur la popularité des documents du système, tel est le cas avec Google qui se base sur l'algorithme RankPage pour définir la pertinence des documents (Page *et al.* 1999), ou elle peut être fondée sur la fréquence des activités de recherche effectuées par les utilisateurs sur le système. Ceci rend la diversité sémantique presque inexistante et ne répond pas à toutes les attentes des différents utilisateurs ou à ceux d'un même utilisateur utilisant le système à différentes périodes de temps.

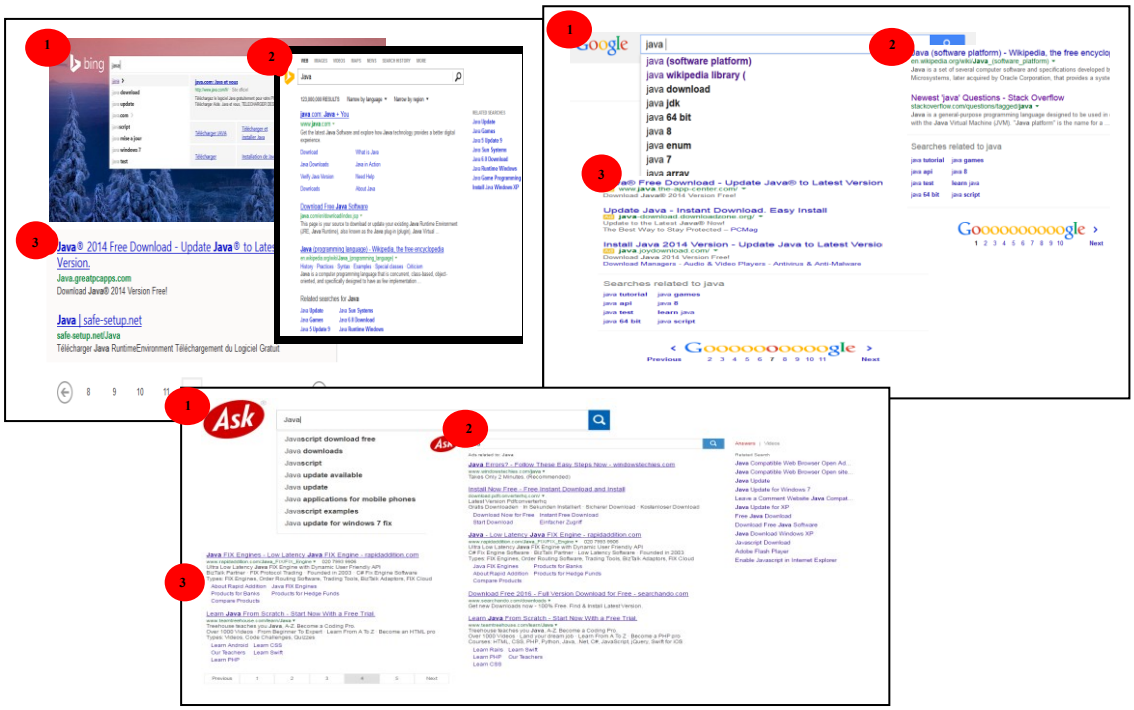


Figure 3. 13. Recherche « Java » avec les moteurs de recherche Google, Bing et Ask.com

Ce problème est atténué dans notre système en se basant sur les différentes interprétations qu'une requête peut avoir dans un dictionnaire linguistique sans porter de préférence sur l'une d'entre elles (cf. figure 3.14). Le système ne fait aucune hypothèse sur les attentes de l'utilisateur. Ainsi, toutes les interprétations qui sont liées au mot Java sont prises en compte dans l'interrogation de l'index documentaire, et les résultats correspondants sont sélectionnés et triés au sein de chaque facette de donnée selon leur score de correspondance BM25F utilisé par Elastic.

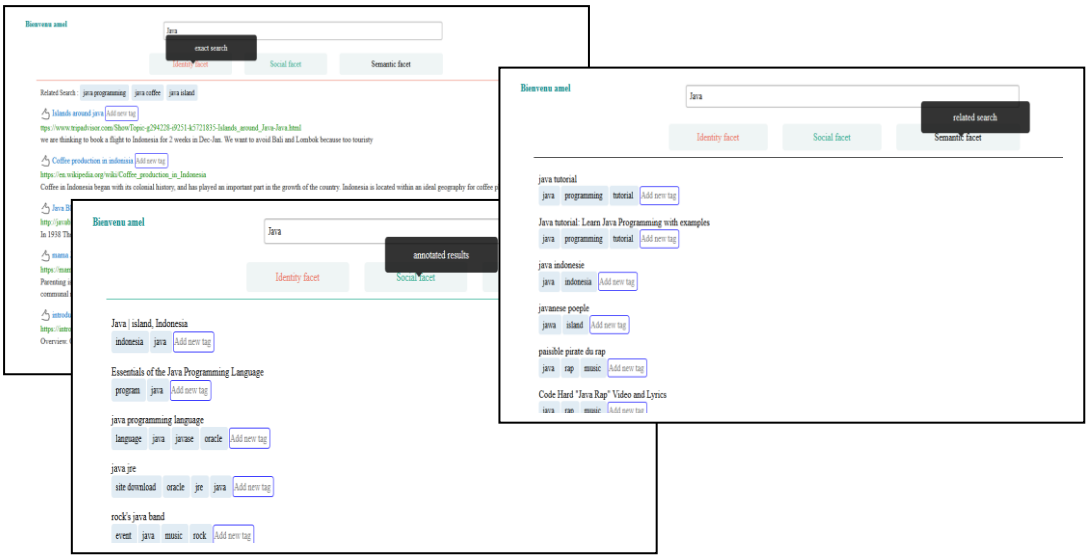


Figure 3. 14. Recherche « Java » avec notre système multi-facettes

Pour cibler la recherche utilisateur à des besoins plus spécifiques, les moteurs populaires précités adoptent des techniques de personnalisation qui s'appuient sur différents facteurs, tels que les historiques des clics et des recherches, l'emplacement géographique, le comportement de l'utilisateur en ligne, etc. Cela peut donner lieu à des différents résultats lorsque la même requête est soumise par différents utilisateurs. En d'autres termes, ces moteurs privilégient certains résultats qu'ils jugent pertinents pour chaque utilisateur. Cette personnalisation est connue sous le nom de la bulle de filtres (Pariser 2011). Elle offre pour chaque utilisateur l'accès à une version différente du web qui lui propose des résultats personnalisés. Par exemple, un utilisateur qui a récemment cliqué sur des sites de kayak en France. Plus tard, quand il recherche des plans de vacances, il pourrait voir des recommandations pour des destinations de vacances en France. Cependant, les besoins de l'utilisateur peuvent changer au fil du temps et devenir différents de ses activités de recherches antérieures (cf. section I.1.4). Ceci engendre une mauvaise interprétation de sa requête. Bien que ces moteurs offrent la possibilité de désactiver cette recherche personnalisée, cela nécessite un minimum de familiarisation avec ces outils de recherche en particulier pour supprimer les historiques des formulaires et des recherches, les cookies, le cache, ou pour désactiver certains paramètres de personnalisation, tels que la localisation géographique, ou encore la déconnexion depuis les sessions qui sont ouvertes sur l'ordinateur et qui sont liées à ces systèmes de recherche, tels que le compte Gmail ou Yahoo pour le cas de Google et Yahoo. En effet, ces paramètres de configuration ne sont pas évidents pour tous les types d'utilisateurs, ce qui peut dégrader la qualité des

résultats qui sont offerts pour ces utilisateurs. En outre, même si un utilisateur est assez familiarisé avec ces outils, l'ajustement de ces paramètres selon ses besoins peut être considéré comme une charge cognitive lourde pour lui, ce qui dégrade l'utilisabilité du système.

Ce problème est atténué dans notre système à travers les facettes de données qui sont proposées à l'utilisateur sur son interface de recherche et l'éventuelle possibilité d'ajouter une nouvelle facette de données qui soit spécifique à chaque utilisateur, et lui laisser le choix entre les vues non personnalisées et la vue personnalisée. Cette vue personnalisée se base sur la définition d'un modèle qui permet d'extraire les intérêts et les préférences de chaque utilisateur et gérer leur évolution au fil du temps pour les intégrer dans le processus de recherche. Cette direction de recherche est présentée dans les deux prochains chapitres 4 et 5.

Un autre problème concerne la façon dont les résultats sont retournés. La majorité des moteurs de recherche populaires retournent les résultats dans une seule vue de données qui englobe une liste ordonnée de tous les documents résultants (cf. figure 3.13 partie 2). L'utilisateur peut se perdre dans une grande masse d'information. En effet, même si l'information recherchée est présente dans la liste des résultats, elle n'est pas toujours facilement accessible par l'utilisateur. Bien que ces moteurs introduisent le paradigme des facettes pour faciliter la recherche et l'interaction avec ces résultats en identifiant différentes catégories pour décrire le contenu et filtrer les résultats, notamment les rubriques: image, vidéo, actualités, cartes, etc. exploitées par Google et Bing. Nous jugeons que cela est loin d'être suffisant puisque les facettes qui sont définies ne sont pas liées à l'aspect sémantique de l'information qui a été appliqué pour décrire et chercher l'information. Ceci n'offre pas à l'utilisateur la possibilité de distinguer entre les différentes interprétations proposées, en particulier avec les requêtes polysémiques.

Ce problème est atténué dans notre système puisque les différentes interprétations d'une requête sont affichées sous forme d'espaces d'interprétation. C'est l'utilisateur qui choisit ce qui répond le mieux à ses besoins. Le système ne fait aucune hypothèse sur ses attentes. Pour cela, différents onglets sont utilisés pour afficher le contenu de ces facettes de données. Par exemple, si un utilisateur fait une recherche sur « tarte aux cerises », le contenu de la facette identitaire comprend les résultats obtenus par un appariement syntaxique entre la requête et les documents dans l'index. Ce sont les documents qui sont décrits dans l'espace identitaire de l'index avec les jetons « tarte » et/ou « cerises » (cf. figure 3.15 Espace-1). Les

résultats de la deuxième facette sociale sont les documents qui ont été annotés par les utilisateurs du système avec un ou plusieurs jetons de cette requête (cf. figure 3.15-Espace2). Les résultats de l'onglet sémantique quant à eux sont les documents qui ont été indexés avec des jetons qui sont liés sémantiquement aux jetons de la requête, tels que les jetons : « gâteau aux cerises », « tourte aux cerises », « galette aux cerises », gâteau aux fruits », « tourte aux fruits » etc. (cf. figure.3.15-Espace3).

Pour mettre en lumière les résultats de recherche, les jetons de la requête ayant correspondu avec chaque document dans l'index, sont mis en évidence sur l'interface de recherche devant chaque résultat. Cela donne à l'utilisateur une certaine transparence sur comment et pourquoi un résultat a été choisi. Chaque espace est également enrichi avec d'autres jetons qui sont liés à la recherche cible. Ces jetons sont considérés comme des valeurs des facettes. Plus concrètement, dans l'espace identitaire les jetons qui cooccurrent avec un ou plusieurs jetons de la requête sont extraits et affichés pour enrichir cet espace de navigation. Par exemple, dans le cas de la requête précédente « tarte aux cerises », les jetons en cooccurrence avec son contenu sont : pâte friable, pâte épaisse, pâte onctueuse, tarte cerises fraiches, etc. (cf. figure 3.15 partie 1). Dans l'espace social, chaque document résultant est renvoyé avec la liste des étiquettes les plus fréquentes qui associent son contenu, telle que par exemple « recette facile », « pâte brisée », « garniture », « cerises », etc. (cf. figure 3.15 partie 2). Des recommandations de recherche sont également proposées dans la facette sémantique, elles sont liées au même sujet de recherche de la requête cible, telles que : « tarte aux cerises et fromage », « tarte aux cerises en canne », « garniture tarte aux cerises », « tarte aux cerises sans gluten », etc. (cf. figure 3.15 partie 3). Ces valeurs de facettes sont offertes sous forme de requêtes prédéfinies sur l'interface offrant à l'utilisateur la possibilité de naviguer/filtrer le contenu du système.

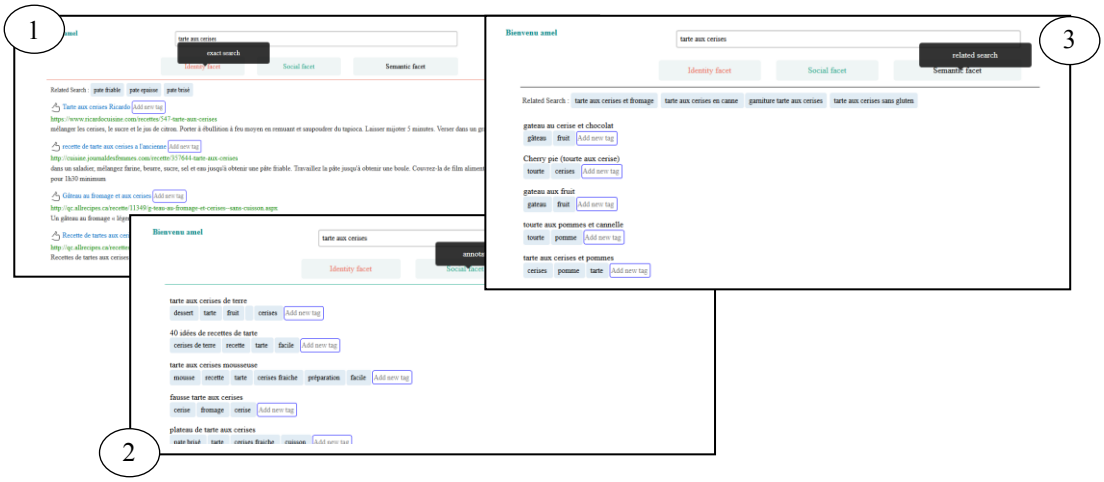


Figure 3. 15. Résultats proposés pour une requête de recherche « tarte aux cerises »

En ce qui concerne l'interprétation de la requête utilisateur, nous avons vu que lorsqu'une requête est composée d'un seul jeton polysémique, le besoin informationnel de l'utilisateur n'est pas clair, ce qui peut engendrer une male interprétation de son contenu par le système. Cependant, lorsque ce jeton polysémique est employé avec d'autres jetons, cela peut aider à désambigüiser son contenu en éliminant les interprétations que le système juge non correspondantes aux attentes de l'utilisateur. Dans certains cas, cela demeure encore difficile, c'est à dire, les jetons qui s'associent au jeton ambigu dans la requête de recherche ne permettent pas au système de cibler le besoin informationnel de l'utilisateur. Nous avons ainsi proposé une technique de désambigüisation pour sélectionner les interprétations les plus similaires au contenu global de la requête. Cette technique évalue la similarité composée de chaque interprétation par rapport au contenu multidimensionnel de la requête puis elle calcule le ou les interprétation(s) les plus similaires en se basant sur le concept du Skyline et de dominance de Pareto (cf. section III.4.12).

Pour illustrer cette technique, prenons l'exemple suivant d'un utilisateur qui soumet la requête « Java animation video ». Comme nous pouvons le voir, cette requête contient le jeton "Java" qui rend la requête ambigüe, et les deux jetons « animation » et « video » ne donnent pas d'indice direct sur les attentes de l'utilisateur, car aucune relation directe de synonymie, d'hyponymie ou d'hyperonymie ne les relie à l'une des interprétations de « Java ». Cela ne permet donc pas au système d'éliminer les interprétations qui ne correspondent pas au contenu global de la requête.

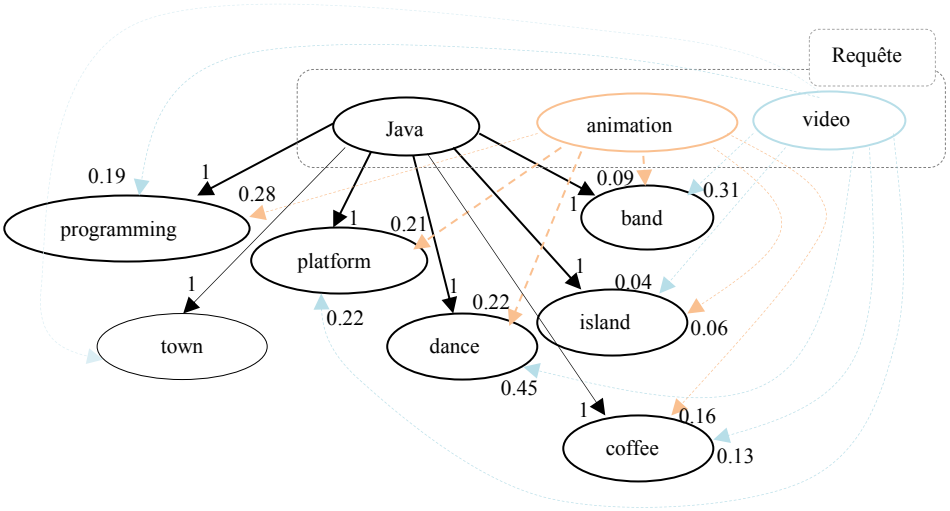


Figure 3. 16. Similarité composée d'une requête de recherche

La technique que nous proposons vise à rapprocher les jetons de la requête aux interprétations possibles du jeton polysémique Java, et cela en évaluant les similarités sémantiques qui existent entre eux dans une ressource sémantique externe.

Jetons de la requête →

Interprétations de Java ↓

$A\ (I, tk_q)$ =

	Java	Animation	Video
I_1 : programming	1	0.28	0.19
I_2 : Platform	1	0.21	0.22
I_3 : Island	1	0.06	0.04
I_4 : band	1	0.09	0.31
I_5 : dance	1	0.22	0.45
I_6: Town	1	0.05	0.04
I_7: Coffee	1	0.16	0.13

Le skyline de A est le sous-ensemble d'interprétations qui ne sont dominées par aucune autre interprétation. Nous pouvons vérifier que le skyline résultant est I_1 = « programming » et I_5 = « dance », car elles ne sont dominées par aucune autre interprétation I_i. Par ailleurs, toutes les autres interprétations sont dominées par I_1. Les deux interprétations résultantes sont exploitées pour enrichir le contenu de la requête. La requête est réécrite comme suit : ((Java and programming) Or (Java and dance)) and

(animation) and (vidéo). Ceci permet d'éliminer les documents qui ne couvrent pas le contexte de la requête.

III.6. Bilan et conclusion

Nous avons présenté dans cette deuxième partie le modèle qui instancie les concepts fondamentaux de notre modèle de RI multidimensionnelle. La principale différence dans ce modèle par rapport à la littérature est la façon dont le système est conçu. Il consiste en une nouvelle méthode d'indexation qui offre une nouvelle façon de récupérer et de présenter les données pour les utilisateurs. Contrairement à ce qui est proposé par les moteurs de recherche classiques, les résultats de recherche de notre système sont proposés de manière plus structurée. L'affichage peut être effectué en fonction des paramètres utilisés dans l'univers d'indexation et des diverses dimensions qui sont préparées pour interpréter la requête de recherche. Cela est réalisé par la distinction des différentes interprétations de cette requête en les séparant en plusieurs instances. Cet affichage est aussi alimenté par des valeurs de facettes dynamiques pouvant être utiles pour attirer l'utilisateur à choisir un résultat ou de filtrer le contenu de la facette. Ces valeurs de facettes sont de différente nature: i) syntaxique en se basant sur la cooccurrence locale des jetons dans les N top documents résultants, ii) sociale en s'appuyant sur les folksonomies du système, iii) sémantique en ayant recours à des ressources prédéfinies telles que les dictionnaires lexicaux, ou par la construction de classes de requêtes à base des sujets de recherche. Ces classes de requêtes sont construites depuis les interactions des utilisateurs sur le système.

Les deux prochains chapitres discutent l'intégration d'une dimension de contenu personnalisé afin de renforcer l'expérience de recherche utilisateur. Il consiste à ne pas limiter l'expérience de navigation aux utilisateurs individuels. Cela est possible en intégrant des techniques de fouille de données pour l'extraction des intérêts et des préférences de ces utilisateurs et des techniques de filtrage collaboratif pour personnaliser leurs résultats de recherche et/ou leur recommander de nouveaux chemins de recherche/navigation. Cela consiste à exploiter les données d'activités de ces utilisateurs pour créer et enrichir leurs profils contribuant à une recherche plus ciblée. Le chapitre prochain présente le modèle de construction d'un profil utilisateur basé sur une représentation multi-niveaux de données d'intérêt.

Chapitre 4 : Profil utilisateur générique basé sur une représentation multi-niveaux de données d'intérêt

IV.1. Introduction

L'objectif de la RI personnalisée est de mieux répondre aux besoins en information des utilisateurs ayant différents objectifs et attentes. Cela est possible en exploitant des paramètres de personnalisation pouvant être intégrés à différents niveaux du système en vue d'influencer positivement le processus de recherche. Ces paramètres identifient les caractéristiques spécifiques de l'utilisateur et définissent son profil. Selon plusieurs chercheurs, un profil reflétant les intérêts récurrents de l'utilisateur représente le facteur le plus important à utiliser pour améliorer de manière significative la performance de la recherche (Gauch *et al.* 2007) (Zemirli *et al.* 2007) (Daoud *et al.* 2008). Dans ce chapitre, nous proposons un modèle générique d'un profil utilisateur qui traduit ses centres d'intérêt recueillis durant ses activités de recherche antérieures (Hannech *et al.* 2016c). Ces centres d'intérêt englobent différentes catégories d'information sémantiques et sociales qui sont représentées en plusieurs niveaux d'abstraction et sont susceptibles d'avoir un impact positif sur le processus de recherche. Une approche d'enrichissement dynamique du contenu de ce profil est ensuite proposée. Elle a pour objectif d'atténuer le problème de manque de diversité des données d'intérêt de l'utilisateur dans son profil. Ce problème se traduit par le fait qu'un utilisateur se voit proposer des suggestions toujours liées aux mêmes centres d'intérêt. Cette approche d'enrichissement se base sur des techniques de fouille de données et de filtrage collaboratif.

Ce chapitre est organisé comme suit. La section 4.2 présente les motivations générales de notre orientation vers une recherche personnalisée, et explique brièvement comment ce type de recherche est utilisé dans notre système. La section 4.3 présente les défis majeurs de la RI personnalisée et les objectifs spécifiques du processus de modélisation du profil utilisateur. Ensuite dans la section 4.4, nous présentons une synthèse de notre modèle du profil utilisateur par rapport aux principales approches discutées dans l'état de l'art. La section 4.5 est consacrée au modèle de construction du profil utilisateur. Elle présente en premier lieu l'analyse comportementale de l'utilisateur à travers le système qui permet de recueillir ses intérêts, la description formelle du système, puis le modèle de représentation et

d'enrichissement du profil. Dans ce modèle, les notations qui sont utilisées dans nos propositions sont d'abord énoncées, puis le processus en question est discuté. Il se compose de deux principales tâches, à savoir la construction des centres d'intérêt de l'utilisateur à base de ses activités de recherche, et l'enrichissement de son contenu à base d'un processus d'inférence collaborative d'intérêts. La dernière section conclut le chapitre.

IV.2. Cadre général et motivation

Suite à une requête de recherche, un système peut retourner un nombre important de résultats qui ne sont pas nécessairement tous pertinents pour l'utilisateur. Ce dernier doit filtrer manuellement les résultats pour écarter ceux qui ne sont pas pertinents pour sa recherche. Afin d'assister l'utilisateur dans ses recherches d'information, nous proposons de filtrer automatiquement ses résultats en tenant compte de son contexte de recherche. Cela consiste à proposer une facette de résultats personnalisés qui soit spécifique à chaque utilisateur. Cette facette répond aux besoins en information de l'utilisateur exprimés par des requêtes de recherche, et à ses caractéristiques contextuelles. Cela est similaire au concept de la bulle de filtres utilisée par les moteurs de recherche populaires (Google et Bing) qui offre à chaque utilisateur une vue personnalisée des résultats des recherches (Pariser 2011). Pour ce faire, un modèle de profil utilisateur est conçu. Il englobe les intérêts et les préférences de l'utilisateur et a pour objectif de réduire son espace de recherche/navigation d'une part, et de rendre exploitables d'une autre part ses données d'intérêt lors de ses recherches. Ces données peuvent être utiles pour améliorer la pertinence des résultats et accélérer le processus de RI. La différence majeure entre notre approche et celles proposées dans la littérature est que les données personnalisées dans notre système sont proposées sous forme d'une facette de données qui s'ajoute à l'espace multidimensionnel discuté dans le chapitre précédent. De cette façon, les résultats de recherche restent diversifiés puisque l'interface de recherche propose plusieurs facettes de données à l'utilisateur qu'il peut explorer selon ses intérêts.

IV. 3. Défis majeurs et objectifs spécifiques

Le domaine de la personnalisation de données est confronté à deux défis majeurs. Le premier défi est lié à la modélisation du profil utilisateur dont fait l'objet ce chapitre. Le second défi se rapporte à

l'exploitation du profil utilisateur conçu. Il définit la technique et le niveau de son intégration dans le processus de RI. Cette étape est discutée à la fois dans ce présent chapitre ainsi que dans le prochain. La finalité de ces deux modèles d'intégration est différente. Le modèle proposé dans ce chapitre vise à offrir à l'utilisateur des données d'intérêt qui servent à enrichir davantage son profil. Ces données représentent de nouvelles expériences de recherche provenant des autres utilisateurs du système, elles ne sont pas directement liées à des besoins spécifiques de l'utilisateur, mais peuvent être utiles et intéressantes pour lui, car elles élargissent les centres d'intérêt couverts sans pour autant noyer l'utilisateur par une grande quantité de résultats. L'objectif du modèle de personnalisation de données discuté dans le prochain chapitre est de filtrer les résultats de l'utilisateur qui sont liés à des besoins spécifiques exprimés par des requêtes de recherche en se basant sur son profil d'intérêts.

D'une manière plus spécifique, les objectifs de la modélisation du profil utilisateur couverte dans ce chapitre s'articulent autour des points suivants :

1) Définition des catégories d'information qui expriment les intérêts de l'utilisateur et qui peuvent être utiles pour améliorer ses recherches.
2) Introduction d'une technique qui aide le système à mieux comprendre les intérêts de l'utilisateur.
3) Définition d'une structure représentative du profil qui permet de s'adapter au changement et à l'enrichissement des intérêts de chaque utilisateur.
4) Adaptation dynamique du profil à l'évolution des intérêts de l'utilisateur au fil du temps.
5) Résolution des problèmes d'ambigüités lexicales des données dans le profil de l'utilisateur, notamment la polysémie et les relations hiérarchiques.
6) Résolution du problème de manque de diversité des intérêts dans le profil de l'utilisateur par l'enrichissement automatique de son contenu. Cet objectif est considéré pour les motivations suivantes : i) lorsqu'un utilisateur a des intérêts limités dans son profil, il ne pourra bénéficier de suggestions qui sont liées à d'autres intérêts. De plus, on remarque de nos jours que les internautes se rendent aux sites commerciaux, tels qu'Amazon et eBay, spécialement pour recevoir les recommandations qui sont liées à certains articles même s'ils n'envisagent pas d'effectuer des commandes. Nous proposons ainsi d'offrir à ces utilisateurs des recommandations qui permettent de réduire leur charge cognitive en prédisant leurs intérêts qui n'ont pas été recherchés/explorés

explicitement (Sugiyama *et al.* 2004) (Song 2014). Par exemple, un utilisateur qui s'intéresse uniquement à des articles de sport peut également recevoir des articles de jeux de vidéo ou d'autres articles connexes. Ces intérêts peuvent l'intéresser même s'il ne les a pas explicitement cherchés. Ce processus est automatique et dynamique dans le sens où il ne nécessite pas l'intervention de l'utilisateur et il s'adapte aux différents besoins des différents utilisateurs et à leurs évolutions.

IV.4. Synthèse

Dans cette section, nous allons discuter les principaux points de comparaison existants entre les méthodologies adoptées dans nos contributions et celles discutées dans l'état de l'art. Cette synthèse permet de positionner notre travail par rapport à la littérature. Les contributions de ce chapitre portent sur la proposition de deux modèles : un modèle de construction et d'enrichissement du profil utilisateur à base de ses activités de recherche individuelles qui couvre les cinq premiers objectifs mentionnés dans la section précédente, et un modèle dynamique d'enrichissement du contenu du profil utilisateur basé sur un processus d'inférence collaborative d'intérêts qui couvre le dernier objectif. La comparaison est alors effectuée selon les différents modèles proposés comme suit :

1. Par rapport au modèle de construction du profil utilisateur à base d'activités de recherche. Ce modèle englobe le processus d'acquisition des données d'intérêt de l'utilisateur et la définition d'une structure adéquate pour leur représentation.

1.1. Détection des intérêts pertinents de l'utilisateur. Pour l'acquisition des données d'intérêt de l'utilisateur, les approches de la littérature s'appuient sur l'analyse de son comportement implicite (nombre de cliques, durée de consultation, etc.) ou explicite (annotation, évaluation par notes ou par commentaire, action de recommandation d'un contenu, etc.) à travers le système de RI. Nous avons vu dans le deuxième chapitre de cette thèse que ces techniques se déclinent en deux catégories selon le contexte de ce système. Ce contexte peut être classique où l'utilisateur est juste un consommateur de contenu, ou il peut être social où le rôle de l'utilisateur passe de consommateur à un producteur de contenu. Dans un contexte classique, les données d'intérêt de l'utilisateur représentent un ensemble de documents qui sont jugés pertinents implicitement ou explicitement par cet utilisateur derrière des requêtes de recherche. Cet ensemble de données est connu sous le nom de l'historique de navigation ou le fichier de journalisation, il

est utilisé par les moteurs de recherche populaires tels que Google, Bing, etc.(Jansen et Spink 2006). Les données d'intérêt de l'utilisateur peuvent être aussi inférées à partir de cet historique de navigation, telles que les domaines d'intérêt de l'utilisateur qui sont liés aux documents de cet historique de navigation (Zemirli 2008) (Daoud *et al.* 2010b) (Hawalah et Fasli 2015), ou les termes les plus fréquents de ses recherches qui sont extraits depuis les requêtes de recherche. Ces termes peuvent être utilisés pour enrichir les requêtes et améliorer les recherches. Cette analyse comportementale souffre parfois du manque de rétroactions explicites de l'utilisateur pour la détection des documents pertinents (Tchuente 2013). De plus, le comportement implicite de l'utilisateur est parfois non fiable pour détecter les réels intérêts de cet utilisateur. Autrement dit, les documents qui sont consultés par cet utilisateur ne sont pas forcément intéressants pour lui (Ma *et al.* 2011) (Mezghani *et al.* 2014). En effet, l'utilisateur ne fournit généralement pas de façon explicite et complète des indications sur ses intérêts.

Dans un contexte social (les réseaux sociaux, blogs, etc.), l'utilisateur est de plus en plus actif, il participe aux discussions et annote des documents. Cela donne plus de transparence sur ses intérêts et permet d'atténuer les problèmes qui sont liés à son comportement classique. Nous soutenons ces propos qui ont été approuvés par plusieurs chercheurs à travers différents travaux de recherche (Beldjoudi *et al.* 2012) (Bouhini *et al.* 2013b) (Mezghani *et al.* 2014). Dans ce type d'applications, les approches se basent sur l'exploitation des étiquettes (Yeung *et al.* 2008) (Meo *et al.* 2013), sur des métadonnées relatives à ces étiquettes (fréquence d'utilisation, pondération, etc.) (De Meo *et al.* 2010), ou sur l'analyse des utilisateurs voisins (dits aussi les utilisateurs proches) qui sont déterminés sur la base de leur comportement social (Cantador *et al.* 2008), etc. Ces éléments sont exploités pour inférer les intérêts de l'utilisateur. Toutefois, l'exploitation de ces informations sociales présente une limitation qui concerne en particulier l'interprétation des étiquettes d'annotation. Ces étiquettes sont parfois ambiguës et personnelles. C'est-à-dire, elles sont spécifiques à l'utilisateur et représentent une description personnelle qui ne peut être compréhensible que par l'utilisateur lui-même. Elles ne fournissement donc pas assez d'informations sur les besoins réels des utilisateurs et n'aident pas le système à détecter les intérêts communs entre les utilisateurs.

Notre approche se base sur une technique hybride qui combine les deux analyses comportementales (classique et sociale) de l'utilisateur. Cette hybridation vise à atténuer les limites qui peuvent être distinguées avec l'une d'entre elles quand elle est adoptée seule. Cela consiste à exploiter

pour déduire les intérêts de l'utilisateur différents objets de contenu qu'il manipule durant ses activités de recherche, en l'occurrence ses requêtes de recherche, ses documents et ses étiquettes. Cette démarche exploite à la fois les deux triplets relationnels « utilisateur-requête-document » (le cas classique) et « utilisateur-étiquette-document » (le cas social) pour extraire et renforcer la relation « utilisateur-document » qui contribue à construire ses intérêts. L'aspect social permet également d'enrichir les liens entre les documents web, et les liens sociaux entre les utilisateurs au sein du système. Il permet aussi d'inférer d'autres relations directes et indirectes entre les différentes entités du système (cf. figure 4.3). En effet, le comportement social ne se limite pas à une simple relation qui relie les documents aux étiquettes d'annotation, mais peut être également exploité pour déduire des comportements similaires entre les utilisateurs. Cela est possible en évaluant les corrélations entre les autres entités du système à qui ces utilisateurs sont liés, notamment les relations entre les documents et les relations entre les étiquettes, tout en intégrant une technique de rapprochement sémantique entre ces entités. Par exemple, deux utilisateurs sont considérés comme similaires s'ils exploitent les mêmes étiquettes ou annotent les mêmes documents, ou s'il existe des relations sémantiques qui relient ces documents les uns aux autres ou relis ces étiquettes les unes aux autres. Cette similarité entre les utilisateurs permet de créer des groupes d'intérêts qui aident à augmenter la portée des profils en s'appuyant sur les intérêts des utilisateurs similaires pour enrichir le profil de chacun (enrichissement collaboratif des profils utilisateurs).

C'est sur la base de toutes ces observations que le comportement hybride de l'utilisateur est adopté pour la construction et l'enrichissement de son profil. À notre connaissance, une telle réconciliation n'est pas encore traitée dans la littérature. Les données d'intérêt de l'utilisateur sont donc représentées et inférées à partir de différents objets de contenu extraits durant ses activités de recherche, ainsi que sur l'inférence collaborative d'intérêts. Le contexte du système est considéré comme hybride.

1.2. Représentation des centres d'intérêt de l'utilisateur. Dans la littérature, les auteurs ont proposé différents modèles qui vont des modèles ensemblistes de données jusqu'aux modèles multidimensionnels plus élaborés intégrant différentes catégories d'information qui sont classées en plusieurs dimensions (cf. section II.2.2.3.3). Dans notre étude, nous proposons un modèle générique qui permet de représenter les données d'intérêt de l'utilisateur en plusieurs niveaux d'abstraction. Il consiste à délimiter les données d'intérêt selon plusieurs aspects qui facilitent leur manipulation par le système de personnalisation et lui

permettent d'appliquer différentes techniques d'analyse selon les besoins de l'application. Ces niveaux vont de la couche concrète qui représente la liste des objets d'intérêt granulaires (dits aussi spécifiques) de l'utilisateur qui sont recueillis durant ses activités de recherche et ses interactions avec le système, en l'occurrence ses facettes d'intérêt, les documents jugés pertinents, les requêtes de recherche, et les étiquettes d'annotation, jusqu'aux couches de haut niveau d'abstraction qui englobent les sujets d'intérêt de l'utilisateur et les groupes de sujets les plus connexes (cf. figure 4.1). Chaque niveau permet de représenter les données d'intérêt de l'utilisateur selon un aspect différent. Cette architecture multi-niveaux permet de regrouper conceptuellement les objets d'intérêt spécifiques de l'utilisateur en des ensembles de domaines d'intérêt connexes qui définissent ses activités de recherche, et de manière sémantique plus élaborée en des sujets de recherche qui regroupent en un seul centre d'intérêt les activités de recherche similaires. Puis, la couche supérieure qui délimite les intérêts de l'utilisateur par groupes de sujets connexes. Ces classifications abstraites aident à résoudre les problèmes qui sont liés à l'ambigüité lexicale de ces objets spécifiques. Elles permettent d'un autre côté d'analyser les similarités existantes entre les utilisateurs du système en les rapprochant à travers les sujets d'intérêt communs, ou par la création de communautés à base des sujets fortement connexes. Ces niveaux d'abstraction sont en particulier utiles lorsque les utilisateurs utilisent différentes requêtes/étiquettes et consomment différents documents qui sont sémantiquement ou contextuellement similaires à travers les couches de représentation supérieures. Il revient donc à rapprocher les entités du système à des niveaux de représentation plus supérieurs.

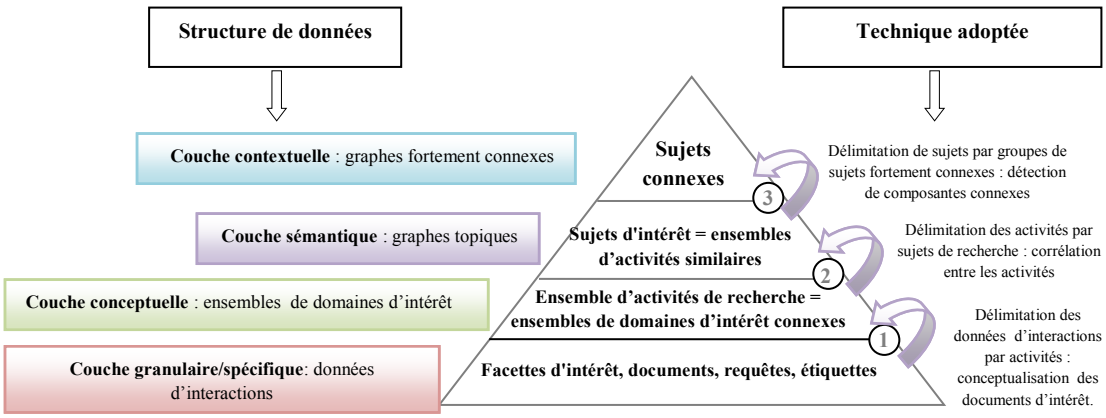


Figure 4. 1. Modèle générique multi-niveaux d'un profil utilisateur

1.3. Adaptation du profil utilisateur. Le profil de l'utilisateur évolue au fil du temps avec l'évolution de ses besoins en information. Dans notre modèle, l'adaptation de ce profil à cette évolution est prise en compte par la mise à jour de son contenu à travers les activités de recherche individuelles de cet utilisateur, et en inférant d'autres intérêts à partir des profils des autres utilisateurs similaires. Cette mise à jour du contenu consiste à archiver (écarter sans supprimer) les intérêts qui sont jugés non pertinents pour l'utilisateur en vue de simplifier leur traitement, et d'en ajouter d'autres qui sont pertinents. Les données ajoutées permettent soit d'enrichir les intérêts récurrents de l'utilisateur ou de considérer de nouveaux intérêts qui n'ont pas été considérés auparavant par cet utilisateur. Cette adaptation est influencée par trois aspects: contextuel, fréquentiel, et temporel.

- Le premier aspect consiste à considérer le contexte des taches de recherche de l'utilisateur dans la construction de son profil. Il s'agit de i) définir des sujets de recherche qui délimitent les différents besoins de l'utilisateur derrière ses activités, et à considérer le changement de ces sujets d'intérêt au fil du temps, et de ii) définir des groupes de sujets qui regroupent les intérêts fortement connexes. Cet aspect contextuel facilite l'analyse des intérêts utilisateur et la gestion de leurs changements au fil du temps.
- L'aspect fréquentiel consiste à alimenter les données d'intérêt de l'utilisateur avec des scores de pondération. Ces pondérations représentent l'importance que l'utilisateur porte pour les données stockées dans son profil. Cette importance considère la fréquence d'utilisation accumulée au fil du temps. Elle permet de définir i) les besoins temporaires et récurrents de l'utilisateur dans son profil, ii) la popularité des objets d'intérêt chez un ou plusieurs utilisateurs. Ces critères d'importance aident à améliorer l'exploitation de ce profil en définissant les préférences de l'utilisateur.
- Au fil du temps les données d'intérêt peuvent fluctuer d'importance chez l'utilisateur même s'ils présentent une fréquence d'utilisation élevée. Par exemple, un nouveau arrivant peut-être intéressé par la location de logement, et une fois qu'il trouve ce qu'il chercher son intérêt va changer. Nous intégrons ainsi l'aspect temporel qui permet d'annoter les données dans le profil de l'utilisateur et de définir leur importance en termes de fraicheur par rapport à une période de temps. Ceci permet de mettre la lumière sur un sous-ensemble d'intérêts qui traduisent les préférences de l'utilisateur dans chaque période de temps indépendamment des deux notions du profil à court et long terme qui sont

employées par plusieurs chercheurs (Gauch, Chaffee et al. 2003) (Zemirli 2008) (Daoud *et al.* 2010b) (Hawalah and Fasli 2015). Celles-ci ne permettent de représenter les données que sous deux angles temporels : le court et le long terme, ce qui ne permet pas de gérer l'évolution des intérêts du point de vue chronologique. Contrairement aux approches qui considèrent que les données d'intérêt récentes des utilisateurs sont les plus pertinentes pour augmenter la pertinence des résultats (Maloof et Michalski 2000) (Zheng et Li 2011; Fu et Kim 2013), ils suppriment donc les informations trop anciennes. Dans notre étude, l'importance est considérée en termes de fraicheur selon n'importe une période de temps qui peut être déterminée par le système ou l'utilisateur. La délimitation temporelle vient donc ajouter un autre filtre au profil de l'utilisateur.

Une technique de combinaison entre ces trois facteurs est proposée.

2. Par rapport au modèle d'enrichissement collaboratif du profil utilisateur. L'enrichissement consiste à mettre à jour les intérêts de l'utilisateur avec de nouvelles données après un traitement prédéfini. Les approches de la littérature adoptent différentes techniques d'inférence d'intérêts qui ont recours en plus du profil utilisateur à d'autres ressources de données sémantiques ou sociales externes (WordNet, Wikipédia, DBPedia, liens web, réseau égocentrique, etc.) (Abdel-Hafez et Xu 2013). Un système d'inférence d'intérêts est plus efficace lorsque les données exploitées durant cette inférence sont pertinentes. Les techniques d'inférence dépendent alors de la qualité des données qui constituent le profil de l'utilisateur et de la technique adoptée pour l'inférence de ces nouvelles données d'intérêt.

Tel que discuté précédemment, les données exploitées dans les approches de la littérature sont déduites soit des activités classiques de l'utilisateur ou de ses activités sociales. De plus, les données manipulées par ces approches appartiennent à un seul niveau de représentation. Dans notre cas, le profil est exploité au sein d'un processus d'inférence d'intérêts qui se base sur l'exploitation des règles d'association extraites à partir du contenu multi-niveaux des profils utilisateurs du système. Comparativement aux approches existantes qui utilisent aussi les règles d'association pour l'enrichissement du profil utilisateur notre approche se distingue par les points suivants :

- L'exploitation d'un comportement hybride des utilisateurs. Il permet d'augmenter la portée des règles d'association sur laquelle se base l'inférence des intérêts de ces utilisateurs. En effet, les profils qui ne contiennent que les informations d'annotation des utilisateurs, tels que les profils qui sont proposés

dans (Schwarzkopf *et al.* 2007) (Zhang et He 2010) (Rebaï *et al.* 2013) (Beldjoudi *et al.* 2016) peuvent être insuffisants pour soutenir la sélection des intérêts pertinents pour ces utilisateurs et la détection des corrélations entre eux. Cette insuffisance peut-être à cause des limitations qui sont liées à ces annotations et qui engendrent une réduction dans le rappel du système. Nous citons par exemple le problème d'incompatibilité de vocabulaire utilisé par les différents utilisateurs pour annoter les mêmes documents.

- La proposition d'une technique qui vise à optimiser le processus d'extraction des données d'intérêt fréquentes des utilisateurs. Cela est effectué en exploitant les niveaux de représentation abstraits de ces données qui sont définis dans les profils des utilisateurs. Il consiste à exploiter à la fois les données d'intérêt spécifiques et génériques, en allant du plus général, les sujets d'intérêt, au plus spécifique, les données d'interactions. Cela permet une sélection progressive des données fréquentes dans les profils des utilisateurs.
- La sélection personnalisée des règles qui sont exploitées pour la prédiction des intérêts de chaque utilisateur. Cela permet d'optimiser l'accès aux règles pertinentes pour chacun d'entre eux.
- La proposition d'une structure à base de sujets de recherche sur laquelle est basée l'extension des données d'enrichissement des profils. Contrairement aux approches qui utilisent des ontologies prédéfinies et spécifiques à des domaines précis pour l'augmentation des données de recommandation (Lu *et al.* 2012) (Adda 2008), la structure proposée dans notre étude est générique, elle englobe tout le vocabulaire utilisé par les utilisateurs. Elle est construite au fur et à mesure que l'utilisateur interagit avec le système.
- La proposition d'une technique de filtrage de données candidates à l'enrichissement du profil de l'utilisateur qui se base sur la similarité entre les utilisateurs. Un document est plus pertinent pour un utilisateur donné lorsqu'il est consommé par un de ses utilisateurs voisins.

2.1. Technique de désambigüisation des étiquettes. L'exploitation des étiquettes sociales au sein du processus d'inférence peut avoir un effet négatif sur le processus de découverte de relations entre les objets de contenu. Cela est dû à plusieurs problèmes qui sont liés à la façon dont les utilisateurs choisissent leurs étiquettes (cf. section I.1.7.). En l'occurrence, nous citons le problème de l'ambigüité polysémique qui se manifeste lorsqu'une étiquette est liée à plusieurs concepts. Cela peut avoir un impact

négatif sur la pertinence des résultats qui sont liés à ces termes. Par exemple, lorsqu'un utilisateur utilise le terme « virus » pour annoter un document, le contenu de ce document peut se référer à un virus informatique ou à un virus humain. Cette exploitation soulève d'importants défis liés à l'extraction des relations sémantiques entre ces termes, en particulier la polysémie qui peut induire à des données non pertinentes et qui réduit la précision du système.

Plusieurs travaux visent à réunir les ontologies et les folksonomies. Cette solution représente une bonne initiative, seulement le problème réside dans la faiblesse des ontologies linguistiques utilisées. Celles-ci fournissent juste un vocabulaire spécifique à un seul domaine de connaissance, ce qui ne permet pas de couvrir tout le vocabulaire utilisé par les différents utilisateurs pour annoter les différents documents. Nous citons aussi la rigidité des représentations ontologiques qui nécessitent l'intervention d'un expert de domaine pour organiser les termes d'annotation sous une forme plus structurée. Ceci est couteux en termes de temps et d'effort. En outre, les termes utilisés par les utilisateurs sont non contrôlés, ils sont parfois personnels et ambigus. Ils peuvent donc être non compréhensibles par une machine ou par un expert de domaine. Ceci limite leur localisation au sein d'une ontologie et limite la découverte des liens sémantiques entre eux. Ces termes sont aussi parfois inadéquats (non représentatifs du contenu des documents auxquels ils sont associés et ne représentant pas les intérêts réels des utilisateurs), ils sont imprécis et ne peuvent pas être exploités seuls pour être reliés automatiquement au contenu d'une ontologie.

D'autres travaux se basent pour la désambigüisation des étiquettes sur des techniques qui s'abstiennent de l'utilisation explicite des ontologies. En l'occurrence, Beldjouh et ses collègues qui considèrent deux ressources étiquetées avec le même terme comme étant similaires lorsqu'elles sont exploitées par des utilisateurs présentant une forte similarité en ce qui concerne leurs historiques d'annotations (Beldjoudi *et al.* 2017). Nous pensons que cette hypothèse n'est pas valable dans toutes les situations, en particulier lorsque l'utilisateur exploite la même étiquette pour annoter plusieurs documents appartenant à différentes interprétations. Par exemple, dans le tableau 4.1 l'utilisateur u_2 s'intéresse à des documents parlant du langage Java, il les annote avec le terme «Java». Cet utilisateur s'intéresse également au domaine de tourisme, il consulte des pages web qui offrent des voyages à l'ile déserte Java et les annotent avec les termes «Java», «Island», «Indonesia». Un autre utilisateur u_3 s'intéresse à la fois à

la cuisine et aux ordinateurs, il annote à la fois les pages web qui proposent des recettes de tarte en pommes et celles qui vendent des ordinateurs de la marque Apple, avec le terme «Apple». Comme nous pouvons le voir, en dépit de la grande similarité de 0.74 qui existe entre les deux utilisateurs u_3 et u_4, les documents annotés avec le terme «Apple» par l'utilisateur u_3 ne peuvent pas être tous proposés à u_4. La même situation s'impose avec les utilisateurs u_1 et u_2 qui présentent une similarité élevée de 0.66. Ainsi, nous pouvons dire que la similarité entre les utilisateurs est loin d'être suffisante pour lever l'ambigüité sur de telles étiquettes. Ces cas ne sont pas bien évidemment fréquents, mais peuvent exister et réduire la précision du système.

Utilisateur	**Étiquette**
u_1	computer, java, programming, Elastic search
u_2	Java, programming, computer, Island, Indonesia.
u_3	Apple, tart, fruit, store, laptop, computer.
u_4	Apple, store, laptop, computer, software
u_5	Apple, computer, virus, chest, flu, headache
u_6	Apple, store, phone, computer, software.

Tableau 4. 1. Ensemble d'un ensemble d'utilisateurs avec leurs étiquettes d'annotation

$\overrightarrow{u_1}$ = ((computer,1), (java,1), (programming,1), (Elasticsearch,1), (Island,0), (Indonisia,0))
$\overrightarrow{u_2}$ = ((computer,1), (java,1), (programming,1), (Elasticsearch,0), (Island,1), (Indonisia,1))

$$\text{Cosinus } (u_1, u_2) = \frac{\overrightarrow{u1}.\overrightarrow{u2}}{\|\overrightarrow{u1}\|^2 * \|\overrightarrow{u1}\|^2} = \frac{(1\,1\,1\,1\,0\,0) * (1\,1\,1\,0\,1\,1))}{\sqrt{(1\;1\;1\,1\,0\,0)} * \sqrt{(1\,1\,1\,0\,1\,1)}} = \frac{3}{\sqrt{4} * \sqrt{5}} = 0.66$$

$\overrightarrow{u_3}$ = ((Apple,1), (tart,1), (fruit,1), (store,1), (laptop,1), (computer,1),(software,0))
$\overrightarrow{u_4}$ = ((Apple,1), (tart,0), (fruit,0), (store,1), (laptop,1), (computer,1),(software,1))

$$\text{Cosinus } (u_3, u_4) = \frac{\overrightarrow{u3}.\overrightarrow{u4}}{\|\overrightarrow{u3}\|^2 * \|\overrightarrow{u4}\|^2} = \frac{(1\,1\,1\,1\,1\,0) * (1\,0\,0\,1\,1\,1))}{\sqrt{(1\;1\;1\,1\,1\,0)} * \sqrt{(1\,0\,0\,1\,1\,1)}} = \frac{4}{\sqrt{6} * \sqrt{5}}\,0.74$$

Le contraire est tout aussi valable, deux documents liés au même terme peuvent être similaires même si les utilisateurs qui les ont annotés n'ont pas une grande similarité. C'est le cas des utilisateurs u_4 et u_5 qui s'intéressent tous les deux à la marque 'Apple'. Donc logiquement les documents qui sont annotés par l'utilisateur u_4 avec ce terme peuvent être proposés à l'utilisateur u_5 et vice versa. Toutefois, avec l'approche proposée dans (Beldjoudi 2015), ces documents ne sont pas considérés comme similaires, car la similarité entre u_4 et u_5 est de 0.36. Même lorsque cette similarité est grande, ce qui est le cas des deux utilisateurs u_4 et u_6 qui s'intéressent tous les deux à la marque « Apple », nous pouvons dire que la similarité entre les utilisateurs qui se base sur la comparaison de leurs étiquettes respectives ou leurs

ensembles de documents, change au fil du temps. Les entités comparées peuvent devenir dissimilaires au fil du temps lorsque l'ensemble des étiquettes augmente ou change, c'est à dire, lorsque les utilisateurs évoluent ou changent d'intérêts (étiquettes ajoutées et/ou supprimées). Cette similarité est donc différente d'un couple d'utilisateurs à l'autre selon les étiquettes qui sont employées par chacun, elle est aussi évolutive au fil du temps.

En nous basant sur toutes ces observations, nous pensons que le contenu des documents est l'indice le plus adéquat pour évaluer la similarité entre deux documents associés à une même étiquette polysémique. Cette adéquation est jugée par la richesse que ce contenu apporte et par le fait que son évolution n'influence pas sur le calcul de similarité entre les documents au fil du temps. Contrairement au contenu imprécis et ambigu des étiquettes, le contenu des documents permet au système de lui associer à un ou plusieurs concepts au sein d'une ontologie de référence. Ces concepts aident à lever l'ambiguïté sur les étiquettes d'annotation et les requêtes de recherche qui sont liées à ces documents. Cela permet également de les relier sémantiquement ou contextuellement en appliquant différentes techniques de rapprochement entre les concepts associés. Aujourd'hui, il existe de nombreuses ontologies topiques qui listent le contenu des documents par catégories et qui peuvent être exploitées pour cette finalité (cf. section V.4.3). Ce contenu documentaire est donc utilisé dans notre étude pour lever l'ambigüité sur de tels objets polysémiques, et pour déduire les intérêts communs entre des utilisateurs au lieu de simples termes d'annotation qui sont généralement employés différemment par les différents utilisateurs.

2.2. Technique d'analyse de variations orthographiques. Nous avons vu dans la section 2.3.2 du chapitre 2 que certaines approches proposent de suggérer aux utilisateurs, lors de l'annotation de leurs ressources, les termes appropriés à leur contenu (Hmimida 2012; Hmimida et Kanawati 2016). Ceci rend le processus d'étiquetage plus contrôlé et peut déplaire aux utilisateurs qui préfèrent employer librement leurs propres termes d'annotation. Néanmoins, dans notre approche, nous proposons une technique d'étiquetage semi-contrôlé pour l'uniformisation des termes ayant la même interprétation tout en laissant à l'utilisateur la liberté quant au choix de ces termes. Cela est effectué par l'analyse des variations qui peuvent exister entre les termes qui sont employés par l'utilisateur lors du processus d'annotation. Ces variations peuvent engendrer une baisse de précision dans le processus de découverte de relations entre les objets. Pour ce faire, lors de l'annotation d'un document, le système exploite la liste des étiquettes qui

ont été utilisées par les autres utilisateurs pour annoter ce même document ou les autres documents qui couvrent le même sujet de recherche, pour proposer à l'utilisateur les étiquettes qui ont une similarité orthographique élevée avec l'étiquette introduite. Une mesure de similarité lexicale entre deux chaines de caractères est exploitée, telle que la distance de Levenshtein (Yujian et Bo 2007) ou celle de Jaro-winkler (Jaro 1989).

De cette façon, nous privilégions l'annotation employée par l'utilisateur au lieu des termes qui sont sélectionnés et proposés par le système. Cela permet d'impliquer l'utilisateur dans l'indexation des documents, ce qui contribue au rapprochement du langage de termes utilisés pour leur indexation à celui utilisé par les utilisateurs lors de la formulation de leurs requêtes de recherche, et contribue à améliorer la pertinence des résultats du SRI.

IV.5. Système de construction du profil utilisateur

IV.5.1. Analyse comportementale de l'utilisateur

Tel qu'il a été discuté dans le chapitre précédent, notre système se base sur le concept de facettes multiples qui permet d'organiser les résultats de recherche en vues de données. Chaque facette est enrichie avec des valeurs de facette que l'utilisateur peut exploiter pour filtrer les résultats ou effectuer d'autres recherches. Le système analyse le comportement de cet utilisateur pour construire et enrichir son profil. Ce comportement peut être classique, il se traduit par la soumission d'une ou plusieurs requêtes de recherche puis l'exploration des résultats en fonctions i) des facettes de données offertes sur l'interface de navigation, ii) des documents retournés dans chaque facette, et iii) des valeurs de facettes qui sont associées au contenu de chaque facette. Ces valeurs de facettes sont considérées comme des requêtes prédéfinies par le système. Le comportement de l'utilisateur peut être aussi un comportement social qui se traduit par l'annotation d'un ou plusieurs documents avec un ou plusieurs étiquettes, ou par l'exploration des documents résultants d'une recherche à travers une ou plusieurs étiquettes qui leur sont déjà associées par d'autres utilisateurs du système. Ces étiquettes sont considérées comme étant des requêtes sociales prédéfinies par le système.

Le système construit le profil de l'utilisateur suite à ses interactions et l'exploite au sein d'un processus de recommandation de données qui permet l'inférence de nouveaux intérêts. Ces intérêts servent d'un côté à enrichir davantage le contenu de ce profil, et d'un autre à assister l'utilisateur lors de ses recherches d'information en lui offrant de nouvelles expériences de recherche. Pour ce faire, la liste de ces recommandations est renvoyée dans une vue de données distincte, l'utilisateur interagit avec les résultats, le système analyse son comportement pour mettre à jour son profil. Ce comportement est résumé dans la figure 4.2 ci-dessous.

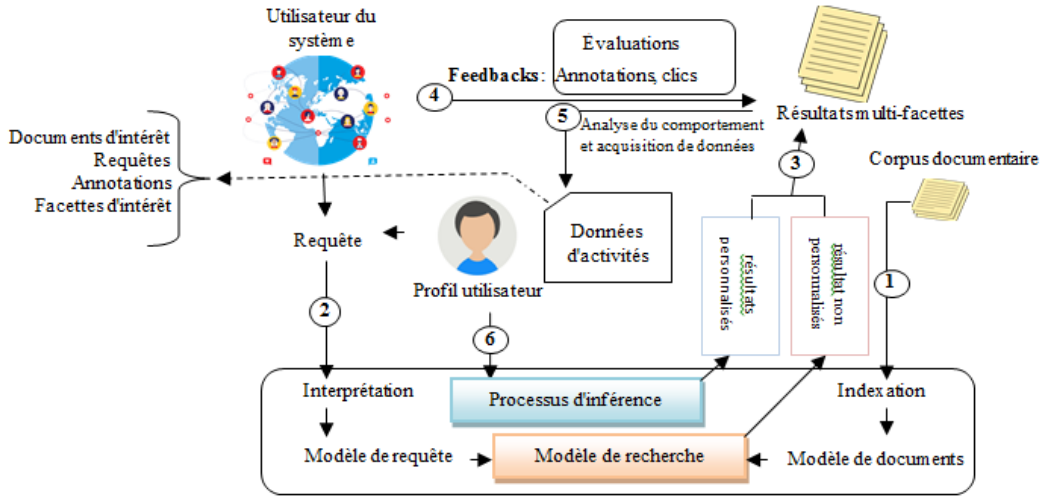


Figure 4. 2. Processus d'acquisition de données d'intérêt de l'utilisateur

IV.5.2. Description formelle d'un système d'analyse de données

Notre système est considéré comme une extension de la notion de folksonomie qui intègre une quatrième dimension représentant les requêtes de recherche des utilisateurs, que nous appelons par Q-folksonomie. Formellement, le système Q-folksonomie est un tuple QF = <U, T, D, Q, R> où U, T, D, Q représentent respectivement l'ensemble des utilisateurs, des étiquettes, des documents et des requêtes de recherche. «R» représente la relation entre un sous-ensemble d'entités précitées du système ($R \subseteq U \times T \times D \times Q$). Dans cette approche, une Q-folksonomie est considérée comme un modèle quadripartite dans lequel les instances peuvent être des documents associés par l'utilisateur à une liste d'étiquettes ou des documents associés par l'utilisateur à une liste de requêtes. Nous extrayons de cette nouvelle extension les

cinq graphes bipartites suivants: «utilisateur-étiquette», «étiquette-document» «utilisateur-requête», «requête-document» et «utilisateur-document». Ces graphes sont représentés par les cinq matrices UT, TD, UQ, QD, et UD qui nous permettent d'analyser plus facilement les corrélations dérivées des différentes interactions utilisateurs.

$$UT_{jk} = \begin{cases} X_{jk} = 1, \text{ Si } \exists\, d \in D, <u_j, t_k, d> \in R \\ X_{jk} = 0, \text{ Sinon} \end{cases} \quad TD_{jk} = \begin{cases} Y_{jk} = 1, \text{ Si } \exists\, u \in U, <u, t_j, d_k> \in R \\ Y_{jk} = 0, \text{ Sinon} \end{cases} \quad QD_{jk} = \begin{cases} F_{jk} = 1, \text{ Si } \exists\, u \in U, <u, q_j, d_k> \in R \\ F_{jk} = 0, \text{ Sinon} \end{cases}$$

$$UQ_{jk} = \begin{cases} Z_{jk} = 1 \text{ Si } \exists\, d \in D, <u_j, q_k, d> \in R \\ Z_{jk} = 0, \text{ Sinon} \end{cases} \quad UD_{jk} = \begin{cases} M_{jk} = 1 \text{ Si } \exists\, t \in T, <u_j, t, d_k> \in R \text{ or } \exists\, q \in Q \\ <u_j, q, d_k> \in R \\ M_{jk} = 0, \text{ Sinon} \end{cases}$$

Tableau 4. 2. Représentation matricielle du système Q-folksonomie

La relation qui relie deux entités dans le système QF peut être de deux types: directe ou indirecte (cf. figure 4.3).

- **Relation directe** : les matrices ci-dessus illustrent les relations directes qui relient les entités du système.
- **Relation indirecte** : deux entités sont indirectement liées si elles sont directement liées à une ou plusieurs autres entités communes. Par exemple, deux utilisateurs sont indirectement liés s'ils sont liés à une ou plusieurs autres étiquettes, requêtes ou documents communs du système. Deux documents sont indirectement liés s'ils ont été annotés avec des étiquettes communes ou par les mêmes utilisateurs.

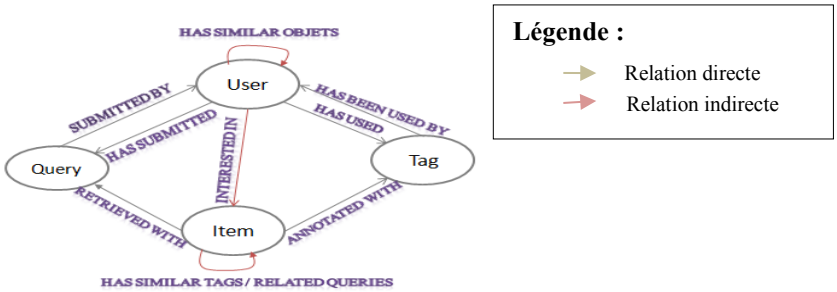


Figure 4. 3. Entités et relations du système Q-folksonomie

IV.4.3. Principaux concepts et notations

Nous présentons dans cette section les principaux notations et concepts qui sont utilisés dans le modèle de construction du profil utilisateur.

Définition 4.1. Ontologie topique. C'est un graphe de concepts qui sont liés sémantiquement les uns aux autres. Chaque concept correspond à un domaine d'intérêt et contient des items (documents Web, vidéos, livres, etc.) dont le contenu se rapporte au domaine correspondant. Ce graphe comporte des composants hiérarchiques créés par des liens de type « est-un » et/ou des composants non hiérarchiques créés par des liens croisés de différents types (Maguitman *et al.* 2005). Il existe plusieurs ontologies qui classifient les objets de contenu en vue de faciliter leur navigation par les utilisateurs. Nous citons des portails en ligne tels que Yahoo, Magellan, Lycos et ODP qui classifient les pages Web par domaines d'intérêt, l'ontologie IMDB qui classifie les films par catégories, etc.

Définition 4.2. Graphe topique. Nous utilisons cette notion pour désigner un graphe de concepts qui traduit un sujet de recherche. Il est constitué d'un ensemble de domaines d'intérêt qui sont extraits et reliés entre eux à travers une ontologie de référence. Chaque nœud dans le graphe contient des items qui représentent les objets d'intérêt de l'utilisateur appartenant à un domaine d'intérêt. En d'autres termes, un graphe topique est considéré comme étant un sous-graphe personnalisé de l'ontologie topique et représente un centre d'intérêt de l'utilisateur, il est par noté par $G = (V, E, Obj)$, où V est la liste des domaines d'intérêt de l'utilisateur qui sont liés un même sujet d'intérêt, E est la liste des relations qui les relie ensemble selon l'ontologie de référence, et « Obj » est l'ensemble des objets d'intérêt de l'utilisateur qui correspondent à ce sujet d'intérêt. Ces objets sont recueillis durant ses activités de recherche (ses interactions avec le système) (cf. figure 4.4).

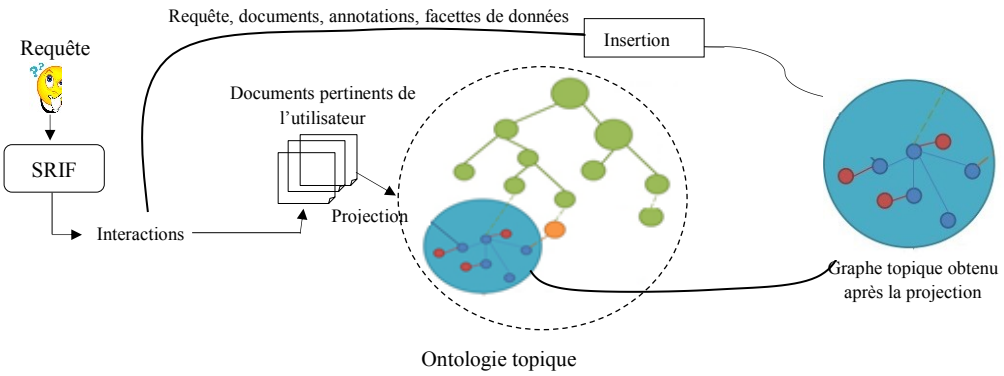


Figure 4. 4. Processus de construction du graphe topique

Définition 4.3. Activité de recherche. Une activité de recherche A^t est l'ensemble d'interactions de l'utilisateur avec le SRI suite à une requête de recherche q^t. L'utilisateur sélectionne et/ou annote un ou plusieurs documents qu'il considère explicitement ou implicitement pertinents. Ces documents sont fusionnés en un seul document D_A, il est représenté sous un vecteur de termes pondérés et noté par V_A. On écrit : $V_A =< (t_1, d_1),.....(t_k, d_k) >$. Une activité de recherche A^t est donc représentée par un ensemble d'objets d'intérêt qui sont recueillis dernière les interactions de l'utilisateur avec un système, notamment les documents qui sont jugés pertinents par l'utilisateur, les étiquettes d'annotation utilisées par cet utilisateur durant cette activité de recherche, ses facettes d'intérêt, et la requête de recherche faisant l'objet de cette activité.

Définition 4.4. Profil de l'activité de recherche. Ce profil traduit le contexte d'une tache de recherche. Il est représenté par un graphe topique pondéré G_A^t constitué d'un ensemble de domaines d'intérêt qui correspond au contenu du document D_A construit durant l'activité de recherche. Chaque domaine d'intérêt est alimenté par un score qui traduit le degré d'intérêt de l'utilisateur pour ce domaine.

Deux requêtes identiques soumises par deux utilisateurs différents ou par l'utilisateur lui-même à deux intervalles de temps différents peuvent avoir deux profils différents de leurs activités de recherche en particulier lorsque cette requête est liée à plusieurs interprétations, par exemple, les requêtes : Java, Apple, virus, Sky, win, etc. L'exploitation de ce profil aide donc à désambigüiser la requête à travers l'analyse comportementale de l'utilisateur. Nous appelons la représentation vectorielle de ce profil par le contexte de la requête, il est noté par $V_{G_A^t}$ tel que $V_{G_A^t} =< (C_1, d_1),.....(C_n, d_n)>$

Définition 4.5. Contexte d'une tache de recherche/sujet de recherche. Deux activités de recherche A^1 et A^2 sont considérées comme étant liées au même sujet de recherche lorsqu'elles partagent plus de domaines d'intérêt dans leurs profils ayant les mêmes rangs d'importance (Tamine *et al.* 2007) (Daoud *et al.* 2010b).

Définition 4.6. Session de recherche utilisateur. Une session de recherche S est un ensemble d'activités de recherche qui sont liées au même sujet de recherche noté par « suj». Elle définit un centre d'intérêt IC_i qui traduit un sujet d'intérêt. Nous notons : S = <$\{A^1, \dots, A^j\}$, suj >. Les activités de recherche formant une session de recherche peuvent être successives ou non successives.

Définition 4.7. Centre d'intérêt. Un centre d'intérêt IC_i, connu aussi sous le nom du profil à court terme (Gauch *et al.* 2003a; Shen *et al.* 2005), est définit dans notre modèle par un graphe topique personnalisé $G_{IC_i}^t$qui englobe les objets d'intérêt de l'utilisateur apparentant à un même sujet de recherche. Ce graphe est le résultat des combinaisons de tous les profils des activités de recherche $\{G_A^1, \ldots . G_A^K\}$ qui sont liées à un même sujet d'intérêt. Cette combinaison peut être évaluée par une opération d'union choisie selon la structure des graphes. Nous notons : $G_{IC_i}^t = \{G_A^1 + \ldots . + G_A^{t-1} + G_A^t\}$

Après avoir donné les principales notations de base qui sont utilisées dans notre modèle, nous passons maintenant à l'approche proposée pour la construction du profil utilisateur.

IV.5.4. Modèle de construction du profil utilisateur

Le processus de construction du profil utilisateur est illustré dans la figure 4.5. Il met en place le processus suivant : l'utilisateur soumet une requête de recherche q^t au système, ce dernier renvoie une liste de résultats parmi lesquels l'utilisateur s'intéresse implicitement ou explicitement à un ensemble de documents notés par D_A^t. Le système exploite cette liste pour créer le profil de l'activité de recherche courante G_A^t. Ce profil représente à la fois le centre d'intérêt courant de l'utilisateur et son profil global P_u^t. Le système traite chaque nouvelle requête afin de créer et enrichir les centres d'intérêt de l'utilisateur en délimitant ses activités par sujet de recherche. Chaque sujet définit un centre d'intérêt IC_i représenté sous un graphe topique G_{Ic_i}. Ce processus de délimitation est basé sur les travaux de (Tamine *et al.* 2007) et (Daoud *et al.* 2010b). Il s'appuie sur l'exploitation d'une mesure de similarité qui permet de calculer la corrélation entre le profil de l'activité courante et le centre d'intérêt courant de l'utilisateur (centre d'intérêt lié à sa précédente requête). Un seuil de corrélation minimal Ω est utilisé, et deux cas sont considérés:

Cas 1. Si la corrélation dépasse la valeur de Ω, nous considérons que la nouvelle requête q^{t+1} est liée au même sujet traité dans la requête précédente. Le système met à jour le centre d'intérêt courant G_{IC}^t avec le profil de l'activité courante G_A^{t+1}.

Case 2. Sinon, l'utilisateur est considéré comme avoir changé de sujet de recherche, et une nouvelle session S^{t+1} commence. Le système vérifie dans ce cas si le nouveau sujet de recherche déclenché derrière la nouvelle requête a déjà été considéré par l'utilisateur dans un temps t' antérieur. Cette analyse est effectuée en corrélant le profil de l'activité courante avec le profil global de l'utilisateur. Celui-ci est

composé d'un ou plusieurs graphes topiques reflétant chacun un centre d'intérêt. Si l'activité courante correspond à un centre d'intérêt dans le profil de l'utilisateur, celui-ci est considéré comme ayant basculé à une session antérieure S'. L'enrichissement du profil se fait en mettant à jour le graphe topique du centre d'intérêt découvert avec le profil de la nouvelle l'activité G_A^{t+1}, et il est considéré comme le nouveau centre d'intérêt courant de l'utilisateur. Dans le cas contraire, un nouveau besoin d'information est considéré et le profil de l'activité de recherche est utilisé pour remplacer le centre d'intérêt courant de l'utilisateur. Le profil de l'utilisateur est exploité comme une base de connaissance pour l'inférence collaborative des intérêts des utilisateurs. Selon ce scénario, le processus de construction du profil utilisateur est décrit par les trois principaux composants suivants :

- La construction du centre d'intérêt associé à un sujet de recherche qui comprend de son côté:
  - La construction du profil de chaque nouvelle activité de recherche.
  - Le mécanisme de détection de changements dans le sujet de la recherche.
- Mise à jour du profil utilisateur selon les deux cas discutés.
- Enrichissement collaboratif du profil basé sur une technique de fouille de données.

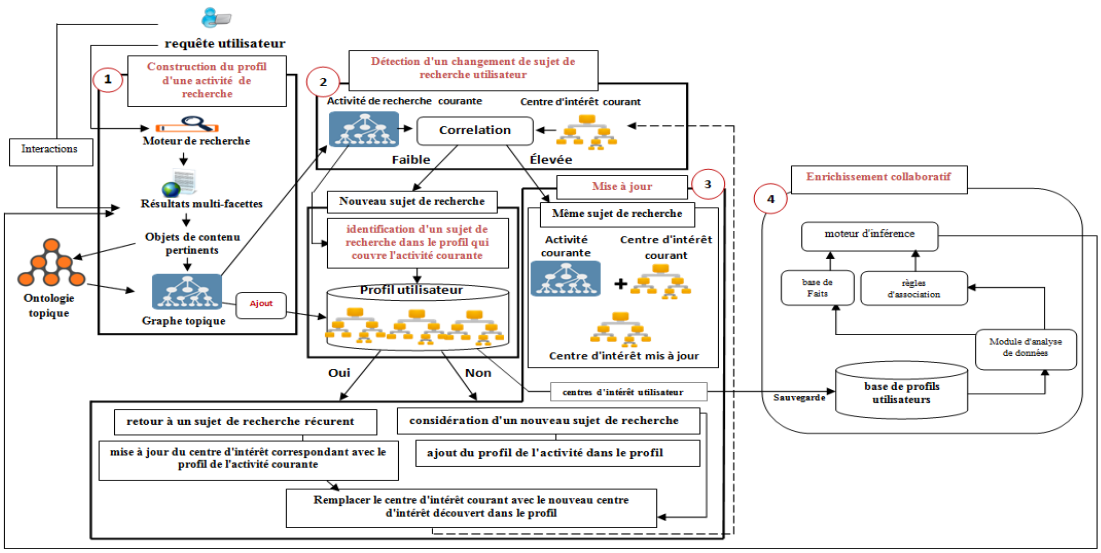


Figure 4. 5. Processus de construction du profil utilisateur

IV.5.4.1. Construction d'un centre d'intérêt utilisateur

Le profil utilisateur est défini par un ensemble de centres d'intérêt. Chacun est déduit lors d'une session de recherche. Un centre d'intérêt est construit en combinant les profils de plusieurs activités de

recherche qui sont liées à un même sujet de recherche. Chacun est représenté par un graphe topique pondéré qui englobe un sous-ensemble de données d'intérêt de l'utilisateur couvrant une sémantique donnée.

Profil d'une activité de recherche. Il reflète le contexte de la tache de recherche de l'utilisateur derrière sa requête. Il est construit à travers le contenu des documents D_A^t qui sont estimés pertinents au cours de cette activité. Pour ce faire, les documents sont projetés sur une ontologie topique afin d'identifier les domaines les plus représentatifs de leur contenu. Cette projection se traduit par le calcul de la similarité cosinus entre le vecteur représentatif de ces documents et chaque vecteur représentant le contenu d'un concept dans l'ontologie de référence. La liste θ des concepts ayant les scores les plus élevés est retenue et utilisée pour créer un graphe conceptuel pondéré. Ce graphe est obtenu en reliant les concepts retenus selon les liens de références de l'ontologie. Les pondérations représentent les scores de similarités obtenus. Puis, en reliant les objets d'intérêt recueillis durant l'activité (étiquettes, documents et la requête de recherche) au graphe conceptuel, le profil de l'activité constitue un graphe topique (cf. figure 4.6). Les objets de contenu sont liés aux concepts ayant une pondération élevée dans le graphe. Nous notons: G_A^t= (θ, E, Ob_A^t). Ce formalisme vient compléter la structure utilisée par d'autres travaux (Tamine *et al.* 2007) (Challam *et al.* 2007) (Daoud *et al.* 2010b) en vue d'adapter l'exploitation des intérêts des utilisateurs aux besoins de notre système de personnalisation.

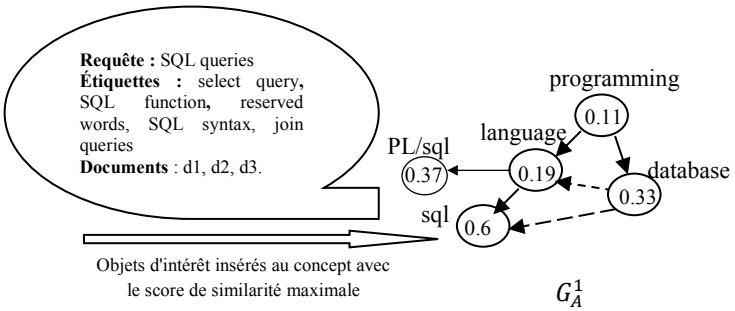


Figure 4. 6. Échantillon du profil d'une activité de recherche liée à une requête « Sql Queries »

Le contexte de la requête est la représentation vectorielle du profil G_A^t , noté par $V_{G_A^1}$. Ce contexte est utilisé pour scruter le changement dans le sujet de recherche de l'utilisateur en corrélant les requêtes successives. Le contexte de la requête « Sql Queries » est représenté par $V_{G_A^1}$, tel que

$V_{G_A^1}$= {(programming, 0.11), (language, 0.19), (database, 0.33), (PL/sql, 0.37), (sql, 0.6)}

Détection du changement dans le sujet de recherche. Chaque nouvelle activité de recherche fait l'objet d'enrichissement du profil utilisateur. Cela est fait par la détection d'un changement dans le sujet de recherche utilisateur en évaluant la similitude entre le profil de l'activité courante et le centre d'intérêt courant de l'utilisateur auquel se rattache la requête précédente q^{t-1}. Cette similitude se base sur la corrélation de Kendall qui compare leurs vecteurs respectifs $V_{G_A^{t+1}}$ et $V_{G_{IC}^t}$. Selon plusieurs chercheurs (Zemirli 2008) (Daoud 2009) cette métrique qui se base sur la corrélation de rangs entre deux variables a donné de meilleurs résultats par rapport à la mesure webJaccard qui se base sur le degré de couverture ensembliste entre ces deux variables. Ces auteurs ont étudié l'influence de deux facteurs sur la détection d'une nouvelle tâche de recherche. Le premier facteur représente l'influence d'un changement de rangs au sein d'un ensemble ordonné d'éléments caractérisant une tache de recherche (ensemble de domaines pondérés, ensemble de termes pondérés, etc.). Le deuxième facteur représente l'influence du degré de couverture ensembliste entre deux ensembles d'éléments relatifs à deux taches de recherche successives. Dans notre cas, la corrélation de Kendall mesure la corrélation entre les concepts d'un centre d'intérêt IC_i et ceux de l'activité de recherche, ce qui nous permet de détecter un changement significatif dans les sujets d'intérêt de l'utilisateur. Ceci est donné par la formule suivante:

$$\text{CorrKendall}\left(V_{G_A^{t+1}}, V_{G_{IC}^t}\right) = \begin{cases} \frac{2(P-Q)}{n\,(n-1)}, \text{ pour } n >= 1 \\ 1, \text{ pour } n=1 \end{cases}$$

Où 'P' est le nombre de paires de concepts concordants (C_i, C_j), 'Q' est le nombre de paires discordantes (C_i, C_j) et 'n' est la dimension des vecteurs à comparer. Une paire de concepts (C_i, C_j) est dite concordante si leurs valeurs de pondération dans les deux vecteurs suivent les équations suivantes: ($x_i > x_j$ alors $y_i > y_j$) ou ($x_i < x_j$ alors $y_i < y_j$). Dans le cas contraire c'est-à-dire lorsque ($x_i > x_j$ alors $y_i < y_j$) ou ($x_i < x_j$ alors $y_i > y_j$), la paire de concepts est considérée comme discordante. Lorsqu'un concept apparaît dans un seul vecteur, il aura un poids nul dans le second.

Si la valeur de corrélation obtenue est plus grande que le seuil prédéfini Ω, la requête est considérée comme étant liée à la même session précédente. Sinon, une nouvelle session est identifiée, dans ce cas, le système vérifie si le nouveau sujet couvre un besoin récurrent dans le profil de l'utilisateur.

Le contexte de cette nouvelle requête est alors corrélé avec chacun des autres centres d'intérêt dans le profil utilisateur. Dans le cas où l'un d'entre eux présente une forte corrélation, la requête est considérée comme étant liée à un besoin d'information antérieur. Un nouveau besoin en information est identifié dans le cas contraire.

IV.5.4.2. Enrichissement du profil utilisateur à base de ses activités de recherche

Le profil de l'utilisateur évolue au fur et à mesure que les interactions de cet utilisateur avec le système évoluent. Il est mis à jour avec le profil de chaque nouvelle activité de recherche A^{t+1}. Cette mise à jour est effectuée selon les deux cas précédemment discutés.

Cas 1: Dans le cas où la requête q^{t+1} est liée au même sujet de la recherche précédente, l'enrichissement du profil utilisateur se fait en enrichissant le centre d'intérêt courant de l'utilisateur avec la nouvelle activité G_A^{t+1}. Cela est effectué en mettant à jour le graphe topique relatif à ce centre d'intérêt avec de nouveaux concepts, liens et objets d'intérêt recueillis durant cette nouvelle activité A^{t+1} .

Cas 2. Dans le cas où la requête q^{t+1}est liée à un nouveau sujet de recherche et qu'aucun centre d'intérêt dans le profil ne lui correspond, le profil de l'activité construit à partir de cette requête est considéré comme étant un nouveau besoin en information avec lequel le profil utilisateur est mis à jour comme suit:

$$P_u^{t+1} = P_u^t \cup \{G_A^{t+1}\}.$$

Dans le cas contraire, le système considère que l'utilisateur a basculé à un sujet de recherche antérieur, le centre d'intérêt identifié est donc mis à jour avec le profil de l'activité de recherche courante. Cette mise à jour est effectuée en combinant leurs graphes topiques G_{IC}^t (V_1, E_1, Obj_1) et G_A^{t+1}(V_2, E_2, Obj_2). Le graphe topique final est noté par G_{IC}^{t+1} (V', E', Obj') tel que $G_{IC}^{t+1} = G_{IC}^t \cup G_A^{t+1}$. L'opération d'union consiste à:

- L'ajout de nouveaux concepts et de nouveaux liens selon la structure de l'ontologie de référence. Cela permet de considérer de nouveaux domaines d'intérêt et de nouveaux objets d'intérêt au sein d'un centre d'intérêt G_{IC}^{t+1} (V', E', Obj'). On écrit :

  $V' = V_1 \cup V_2$ et $E' = E_1 \cup E_2$ tel que $\forall\, v_i \in V_1 \cap V_2$, $Obj_{G_{IC}^{t+1}}(v_i) = Obj_{G_{IC}^t}(v_i) \cup Obj_{G_A^{t+1}}(v_i)$

- Chaque concept dans le graphe G_{IC}^t est également mis à jour avec un nouveau score de pondération si celui-ci est un domaine commun entre les deux graphes G_A^{t+1} and G_{IC}^t. Le score de pondération final

noté par $score_{G_{IC}^{t+1}}(c_i)$ correspond à la moyenne arithmétique des deux valeurs de pondération $score_{G_{IC}^{t}}(c_i)$ et $score_{G_{A}^{t+1}}(c_i)$ dans les deux graphes G_{IC}^{t} et G_{A}^{t+1}.

- La moyenne arithmétique de tous les scores de pondération d'un centre d'intérêt représente le degré d'intérêt de l'utilisateur pour le sujet d'intérêt. Ce score d'intérêt est noté par score (suj_i).

De cette façon, un profil d'utilisateur P_u^t se compose d'un ensemble de m-graphes topiques pondérés, tel que : $P_u^t = \{G_{IC_1}^t,\ldots,G_{IC_m}^t\}$, chacun représente un centre d'intérêt qui regroupe en un cluster sémantique un sous-ensemble de données d'intérêt de l'utilisateur appartenant à un même sujet de recherche. Ces clusters sémantiques aident à désambigüiser le contenu de ces données lors de la recherche d'information.

Enrichissement temporel des données d'intérêt: Chaque objet d'intérêt dans le profil d'un utilisateur u est annoté temporellement avec une date de consommation $date_{obj_i}$ qui permet d'inférer sa fraîcheur. La fraîcheur d'un document d_i est calculée par rapport à une période de temps Δt comme suit :

$$F(d_i, \Delta t, u) = \begin{cases} \alpha\left(\frac{\sum_{j=1}^{n} F(t_j, \Delta t, u)}{n}\right) + \beta * \frac{1}{SP_{d_i}}, & \text{si } date_{d_i} \in \Delta t \\ 0, & \text{sinon} \end{cases}$$

Où $F(t_j, \Delta t, u)$ est la fraîcheur d'une étiquette t_j associée au document d_i. Elle est égale à l'inverse du temps écoulé depuis sa dernière exploitation par l'utilisateur pour annoter le document cible d_i jusqu'à la fin la période Δt considérée. Donc, plus cette période de temps écoulé est petite, plus la valeur $F(t_j, \Delta t, u)$ est grande. «n» est le nombre d'étiquettes qui sont associées à d_i. L'évaluation de la moyenne arithmétique des fraicheurs relatives aux étiquettes du document permet pour donner plus d'importance aux documents qui ont été annotés plusieurs fois par l'utilisateur dans la période Δt.

SP_{d_i} est le temps écoulé depuis la dernière date de consommation de ce document par l'utilisateur jusqu'à la fin de la période Δt considérée. Ainsi, plus cette période écoulée est petite, plus le document est considéré comme plus frais. Lorsque la date de consommation du document n'est pas incluse dans la période Δt, celui-ci ne correspond pas à la qualité de fraicheur ($F(d_i, \Delta t, u) = 0$), il est donc éliminé.

Les deux coefficients de pondération α et β sont des constantes qui reflètent le degré d'influence de chaque paramètre considéré dans l'évaluation de cette fraicheur. Cela permet de donner plus d'importance à un comportement social de l'utilisateur par rapport à son comportement classique, c'est- à-dire, la fraicheur du document est plus importante lorsque le document est consommé par l'utilisateur de manière

sociale à travers un ou plusieurs étiquettes que lorsqu'il est consommé de manière classique à travers des clics.

Pour évaluer cette fraicheur, les d'intérêt des utilisateurs sont délimités par périodes de temps. Cela permet au système d'appliquer des filtres qui offrent à l'utilisateur la possibilité de personnaliser ses résultats selon une préférence de temps. La délimitation temporelle a été utilisée dans (Mezghani *et al.* 2014) (Boudiba et Ahmed-Ouamer 2017) sur un profil social constitué d'un ensemble d'annotations en vue d'adapter le contenu social de l'utilisateur à ses annotations récentes. Dans notre cas, cette délimitation temporelle permet d'appliquer un filtre supplémentaire aux centres d'intérêt de l'utilisateur pour mieux gérer leur exploitation au sein du processus de RI. Pour ce faire, plusieurs types de préférences peuvent être définis pour déterminer la période Δt que l'utilisateur peut utiliser en tant qu'un filtre avec sa requête de recherche :

Légende :
1 : méthode d'identification interactive des préférences utilisateur
2 : méthode d'identification automatique des préférences utilisateur

	Préférence Δt	**Signification**	**Fraicheur**
1	**Fi1re 1**: Since $date_i$	Lorsque l'utilisateur souhaite recevoir des résultats selon les activités les plus récentes depuis un instant $date_i$. Ce filtre considère uniquement les intérêts depuis $date_i$ jusqu'à l'instant présent.	$F(d_i, \Delta t, u) = \begin{cases} 1, \text{ si } date_{d_i} \in \Delta t \\ 0, \text{ sinon} \end{cases}$
	Filtre 2: Between $date_1$ and $date_2$	Lorsque l'utilisateur souhaite personnaliser ses résultats selon les activités de recherche qui ont été effectuées entre deux instants $date_1$ et $date_2$. Ce filtre considère les intérêts appartenant à l'intervalle défini.	
	Filtre 3 : Until $date_i$	Lorsque l'utilisateur souhaite archiver ses activités récentes sans à devoir les supprimer. Ce filtre permet de considérer son historique jusqu'à $date_i$.	
2	Aucun filtre	Si aucun filtre n'est appliqué, la fraicheur d'un objet est calculée à travers l'équation suivante :	$\alpha(\frac{\sum_{j=1}^{n} F(t_j, \Delta t)}{n}) + \beta * \frac{1}{SP_{\Delta t}}$

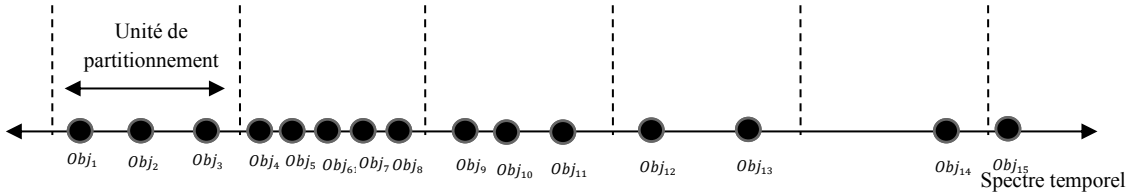
Tableau 4. 3. Préférences de l'utilisateur pour la fraicheur de ses données d'intérêt

Lorsqu'aucun filtre n'est appliqué par l'utilisateur, la période de temps Δt est déterminée depuis la dernière date de consommation de l'objet jusqu'à l'instant présent de la personnalisation (cf. tableau 4.3 partie 2).
La recherche personnalisée à base de filtres est une méthode interactive qui peut être combinée à une méthode automatique d'identification des préférences de l'utilisateur, pour améliorer la RIP. Elle permet à l'utilisateur d'exprimer explicitement ses préférences en terme temporel. L'utilisateur représente alors l'agent qui aide le système à désambiguïser sa requête et écarter les informations non pertinentes qui ne répondent pas à ses attentes. Nous citons l'exemple d'un utilisateur qui s'est intéressé aux virus

informatiques et aux virus humains, lorsqu'il soumet à nouveau une requête en relation avec les virus, il peut indiquer au système le contexte de sa requête en appliquant un filtre sur le contenu de son profil.

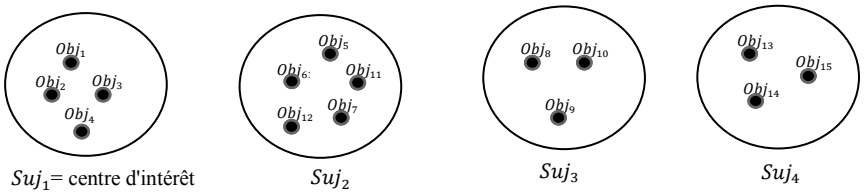
La figure 4.7 illustre différentes techniques de délimitation du contenu du profil utilisateur. Cette illustration permet de montrer l'adaptabilité de notre modèle de représentation de ce profil à différentes techniques de recherche.

Cas 1-Avec uniquement une délimitation temporelle (Mezghani *et al.* 2014) (Boudiba et Ahmed-Ouamer 2017)



L'unité de partitionnement temporel peut être définie en termes de n'importe quelle unité temporelle : minute, heur, jour, semaine, mois, ou année.

Cas 2-Avec délimitation sémantique en sujets d'intérêt (Daoud *et al.* 2010b) :



Cas 3-Avec délimitation sémantique et temporelle (Hannech *et al.* 2016c) :

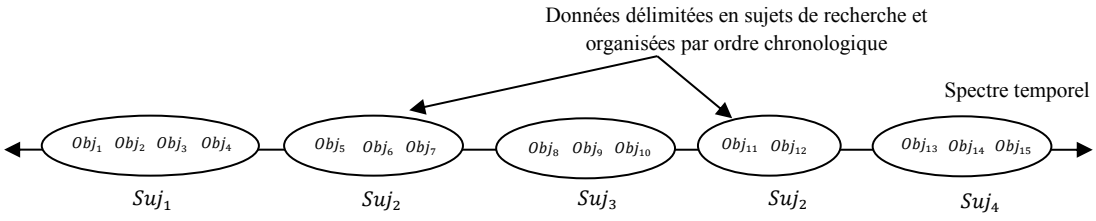


Figure 4. 7. Différentes techniques de délimitation du contenu du profil utilisateur

Prenant l'exemple d'un utilisateur qui soumet une requête et souhaite recevoir des résultats personnalisés qui soient liés à ses activités récentes effectuées durant les 3 derniers jours. Il applique alors le filtre Filtre 1= « Since 3 days ». En identifiant le contexte de cette tache de recherche et le projetant sur le contenu du profil utilisateur, le contenu du premier profil (cas 1) n'est pas adaptable à une telle recherche

puisque les données sont délimitées uniquement par ordre chronologique. Dans le deuxième cas, même si le système est capable d'identifier les centres d'intérêt qui couvrent le contexte de la recherche, il est impossible d'identifier les données d'intérêt qui répondent aux besoins de l'utilisateur en terme temporel. Tandis qu'avec la structure proposée dans notre modèle (cas 3), le système applique le filtre temporel pour sélectionner les objets d'intérêt des 3 derniers jours, puis identifie parmi les centres d'intérêt qui sont liés aux objets sélectionnés celui ou ceux qui couvrent le contexte de la requête.

Cette délimitation hybride permet alors au système d'appliquer différentes techniques de combinaison entre le contexte et la fraicheur des données pour mieux gérer automatiquement leur exploitation au sein du processus de RIP.

Délimitation contextuelle de données d'intérêt des utilisateurs. Cette délimitation consiste en la création d'un niveau d'abstraction plus général que les sujets d'intérêt qui permet de regrouper contextuellement les données d'intérêt des utilisateurs. Cela est effectué en localisant les groupes de sujets fortement connexes (SFC). Ce processus est réalisé en deux étapes : i) la construction d'un graphe général G_g qui englobe tous les centres d'intérêt des utilisateurs en une seule structure, où les nœuds représentent les sujets d'intérêt et les arcs sont les relations qui existent entre eux. ii) Puis, la détection des composantes fortement connexes (CFC) au sein du graphe G_g obtenu. L'orientation des arcs dans le graphe $G_g$est importante dans notre cas. Elle représente le degré de connectivité entre les sujets d'intérêt qui permet la détection de SFC au sein de ce graphe. Un arc multidirectionnel permet de représenter une forte connectivité entre deux sujet d'intérêt et un arc unidirectionnel représente une connectivité d'un seul côté à travers un ou plusieurs lien de références qui relient les deux sujets dans l'ontologie de référence.

Donc, l'idée est rapprocher deux par deux les centres d'intérêt des utilisateurs par l'évaluation des recouvrements structurels entre leurs graphes conceptuels respectifs. Pour évaluer ce recouvrement, nous nous basons sur la mesure de distance composée proposée dans (Fernández et Valiente 2001). Elle est basée sur la combinaison du plus grand sous-graphe commun de deux graphes (MCS) et leur plus petit super-graphe (mcs). Ce choix se justifie par le fait que le MCS de deux graphes permet de mesurer la similarité entre eux à un niveau de généralité basé sur les niveaux supérieurs dans l'ontologie de référence, et le mcs permet de mesurer cette similarité à un niveau de spécificité basé sur les niveaux les plus inférieurs de cette ontologie. Plusieurs cas de recouvrement sont possibles :

- **Cas 1**: recouvrement syntaxique entre un sous-ensemble de concepts Il est évalué par le nombre de concepts communs entre deux graphes g_1 et g_2. Si la similarité est élevée, les deux graphes sont considérés comme étant fortement liés. Cela résulte par la création d'un arc bidirectionnel entre les deux nœuds g_1 et g_2 dans G_g.
- **Cas 2**: lorsqu'aucun recouvrement conceptuel n'existe entre les deux graphes, et la présence d'un ou plusieurs liens de référence entre leurs concepts dans l'ontologie de référence. Cela résulte par la création d'un ou plusieurs arc entre les nœuds g_1 et g_2 dans G_g de la même direction que les liens de référence qui existent entre les deux graphes dans cette ontologie de référence.
- **Cas 3**: pas de concepts communs ou de liens de référence qui relient les deux graphes, donc pas de lien direct entre les deux nœuds g_1 et g_2 dans G_g.

Le résultat est un graphe G_g qui relit les sujets en une seule structure par similarité structurelle. La détection de CFC au sein de ce graphe fait l'objet de localisation de SFC. Pour cela nous appliquons l'algorithme de Kosaraju (Callahan et Kosaraju 1993) qui se base sur la recherche de parcours en profondeur (ou DFS, pour Depth First Search). Il se base sur les principes suivants :

- Un graphe orienté est fortement connexe s'il existe un chemin entre toutes les paires de nœuds.
- Une composante fortement connexe (CFC) d'un graphe orienté est un sous-graphe maximal fortement connecté. Un sous-graphe est considéré maximal lorsque le nombre de nœuds qui le constituent est maximal. Par exemple, il existe 4 CFC dans le graphe G_g de la figure 4.8.

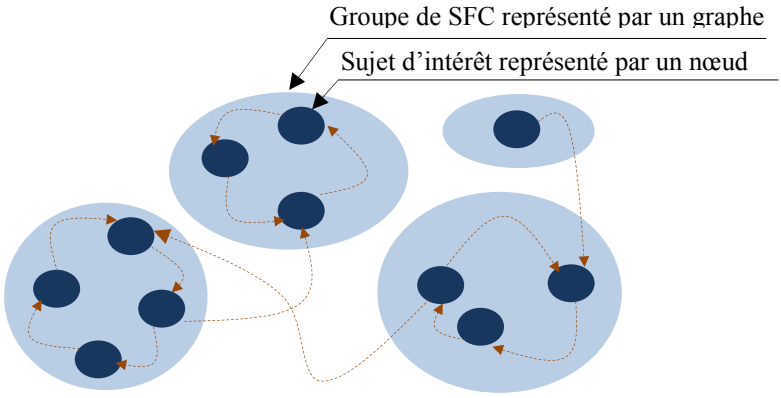


Figure 4. 8. Exemple d'un ensemble de groupes de SFC

Le résultat est un ensemble de groupes de SFC. Les objets d'intérêts appartenant à un même groupe constituent un même cluster contextuel.

Enrichissement du système Q-folksonomie. Le processus de délimitation des données d'intérêt par sujets de recherche et par sujets fortement connexes permet d'enrichir le système avec de nouvelles dimensions abstraites notées SD (pour Semantic Dimension) et CD (pour Contextual Dimension). On écrit QF = <U, T, D, Q, SC, CD, R>. Cet enrichissement permet la création de nouveaux graphes bipartis et tripartis qui relient les entités U, T', D et Q à ces nouvelles dimensions qui facilitent leur analyse dans le système, en particulier les graphes : (Utilisateur-Tag-sujet) (Utilisateur-Tag-CD). (Utilisateur-requête-sujet), (Utilisateur-requête-Contexte) (Utilisateur-document-sujet), (Utilisateur-document-contexte), (requête-Contexte), (requête-sujet), (étiquette-contexte) (étiquette-sujet), (document-contexte), (document-sujet). Ces relations contribuent au rapprochement sémantique et contextuel des entités du système U, T, D, Q et aide à :

- La désambigüisation du contenu polysémique des objets d'intérêt des utilisateurs lors du processus d'inférence d'intérêts à travers leurs sujets respectifs. Cette inférence s'appuie sur les relations qui relient les entités du système les uns aux autres. Cette désambigüisation promet donc une meilleure recherche et inférence d'intérêts.
- La création de communautés d'intérêts à travers les groupes de sujets connexes (les classes contextuelles de données).

IV.5.4.3. Illustration du processus de construction du profil utilisateur et son évolution à travers les activités de recherche

Nous avons choisi d'utiliser le répertoire du web ODP (Open Directory Project), connu aussi sous le nom de DMOZ, comme ontologie de référence conceptuelle compte tenu la quantité des documents web et la diversité des catégories qu'elle contient (environ 1 million de catégories et 4 millions de pages web). Cette ontologie inclut trois types de relation: 1) la relation « is-a » notée par « T » qui classifie hiérarchiquement les concepts du plus général au plus spécifique. 2) la relation « symbolic » notée par « S » qui permet de relier les concepts dans deux hiérarchies différentes. Et 3) la relation « related to » noté par « R » permet d'aller vers des concepts traitant le même concept sans partager des objets en commun. La diversité des concepts et de relations offerte par cette ontologie de référence améliore la détection de domaines d'intérêt relatifs aux activités de recherche des utilisateurs. Ceci permet d'avoir une représentation riche de leurs profils (les profils des activités de recherche) qui peut à son tour

aider à améliorer la détection de corrélations entre les activités de recherche et la détection de similarités entre les sujets connexes.

Pour illustrer le processus de construction du profil utilisateur, considérons l'exemple d'un utilisateur qui a effectué les recherches suivantes : Q = {q_1= sql language, q_2= PL/SQL, q_3= java programming}. Pour chaque itération dans la liste Q nous présentons l'état du profil utilisateur.

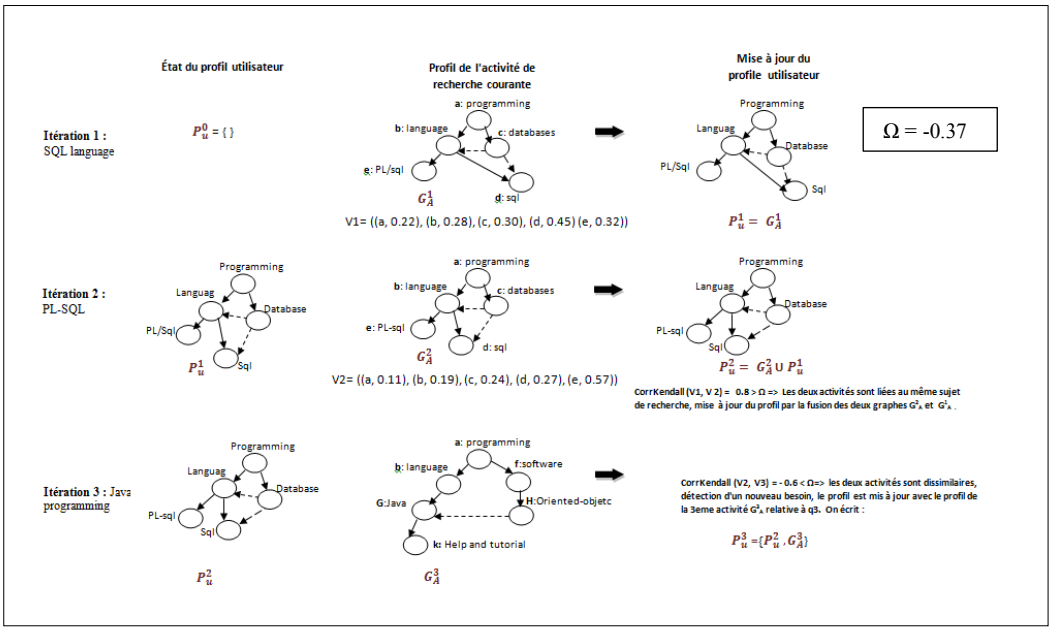


Figure 4. 9. Exemple illustratif de l'évolution du profil utilisateur

IV.5.4.4. Enrichissement du profil utilisateur à base d'un processus d'inférence collaborative de données d'intérêt : recommandation hybride basée sur l'exploitation des règles d'association

Règles d'association et Q-folksonomie. Nous avons vu dans le deuxième chapitre que les règles d'association ont été utilisées pour découvrir des relations intéressantes entre les variables dans de grandes bases de données. Il s'agit de déterminer les items qui apparaissaient fréquemment ensemble. Nous parlons dans ce cas d'itemsets fréquents. Par exemple, dans les données transactionnelles des points de vente, si le pain et le lait sont présents dans 35% des chariots d'achat (on écrit : si pain alors lait), cela indique que si un client a acheté du pain, il est susceptible d'acheter du lait. Dans ce cas, « Pain » est appelé antécédent de la règle, et « lait » est le conséquent.

Nous nous basons sur ce concept pour mettre en évidence les données d'activités des utilisateurs sous la forme de règles d'association. Elles sont exploitées pour fournir à un utilisateur cible de nouvelles expériences de recherches basées sur des expériences des autres utilisateurs présentant une similarité comportementale avec lui. Nous avons vu dans la section II.2.3.1.5 que l'extraction des règles d'association passe généralement par trois principales étapes: la préparation de données, l'extraction des itemsets fréquents et l'extraction des règles d'association. Dans notre système, ce processus suit les étapes suivantes :

- **Etape1-Sélection et préparation de données.** Nous nous basons sur la délimitation contextuelle de données pour préparer les intérêts des utilisateurs qui servent à extraire nos règles d'association. Ces données d'intérêt sont donc partitionnées en plusieurs classes contextuelles. Chacune constitue un ensemble de sujets fortement connexes (SFC). Cette répartition est fondée sur l'hypothèse que les sujets les plus connexes sont ceux qui contiennent généralement les itemsets fréquents. Cette répartition contribue à la réduction de la quantité des itemsets qui peuvent être extraits des items de départ et sur lesquels se base la sélection des itemsets fréquents dans la prochaine étape. Si on suppose que N items de départ existent, nous avons (2^N-1) itemsets extraits (cf. section figure 2.7). Lorsque l'ensemble des items est réparti en plusieurs groupes N_1, N_2, … N_k (tel que $N = N_1+N_2+\ldots+N_k$), le nombre d'itemsets qui sont extraits de ces groupes est réduit à $(2^{N1} + 2^{N2}+\ldots.+2^{Nk})$-k. Ceci est considérablement plus petit que $(2^N-1) = (2^{(N1+N2+..+\ Nk)})-1= (2^{N1} \times 2^{N2}\times\ldots.\times 2^{Nk})-1$. Cette réduction aide alors à réduire la complexité de l'étape suivante. Celle-ci consiste à extraire les itemsets fréquents en faisant correspondre chaque itemsets à chaque transaction de la base (dans notre cas il s'agit du nombre des utilisateurs du système). La complexité de cette étape est de O(L*M) où « L » est le nombre d'itemsets extraits de cette étape, et M est le nombre de transactions qui sont considérées dans cette extraction. Ainsi lorsque la valeur de « L » est réduite, la complexité de cette étape est à son tour allégée.
- **Étape 2-Extraction des sujets les plus fréquents (appelés les itemsets génériques fréquents) au sein des classes contextuelles:** ce processus s'effectue en partant du niveau plus général, les sujets d'intérêt, au plus spécifique, les objets d'intérêt granulaires (spécifiques). Ceci est basé sur l'hypothèse que i) si un sujet de recherche est non fréquent alors il possède des objets d'intérêt granulaires non fréquents et ii) si un k-itemset générique (un ensemble de k sujets) n'est pas fréquent

donc aucun k-itemset spécifique construit par combinaison des objets spécifiques de ces k sujets, n'est fréquent. L'application de ces deux hypothèses permet d'éliminer les objets d'intérêt granulaires non fréquents en se basant sur leur plus haut niveau d'abstraction. Ceci réduit le nombre d'itemsets qui peuvent être extraits et peut donc être utile pour optimiser davantage le temps de calcul. Pour l'accomplissement de cette étape, pour chaque classe contextuelle préparée, les données transactionnelles des utilisateurs sont préparées. Il consiste à représenter chaque utilisateur par l'ensemble de ses sujets d'intérêt appartenant à la classe contextuelle en question, puis les ensembles de sujets les plus fréquents sont extraits de cette classe. Le résultat est un ensemble de classes où chacune englobe les sujets des utilisateurs les plus fréquents appartenant à un même groupe de sujets connexes. Chaque sujet est lié à l'ensemble des objets d'intérêt granulaires de ces utilisateurs.

- **Étape 3-Extraction des objets d'intérêt (itemsets spécifiques) les plus fréquents.** De la même manière, les objets d'intérêt spécifiques les plus fréquents sont extraits de chaque classe de sujets fréquents obtenue de l'étape précédente. Pour ce faire, dans chaque classe, nous représentons chaque utilisateur dans le système par une transaction de ses objets d'intérêt granulaires (ses requêtes et étiquettes) qui sont liés aux sujets fréquents de la classe, puis les k-itemsets spécifiques sont extraits. Le résultat final est un ensemble de classes, chacune englobe les itemsets spécifiques les plus fréquents, appartenant à un même groupe de sujets connexes.

  Dans les deux étapes 2 et 3, la fréquence d'un k-itemset (spécifique ou générique) est calculée à travers la métrique de support minimum (cf. définition II.1).

- **Étape 4-Extraction des règles d'association.** En appliquant l'algorithme d'exploration de données Apriori (Agrawal *et al.* 1993), les règles d'association intéressantes sont extraites à partir des itemsets fréquents obtenus dans chaque classe contextuelle. Ainsi, les règles obtenues sont délimitées à leur tour par classes où chacune est destinée à représenter les corrélations d'une communauté d'intérêt construite à base de sujets fortement connexes.

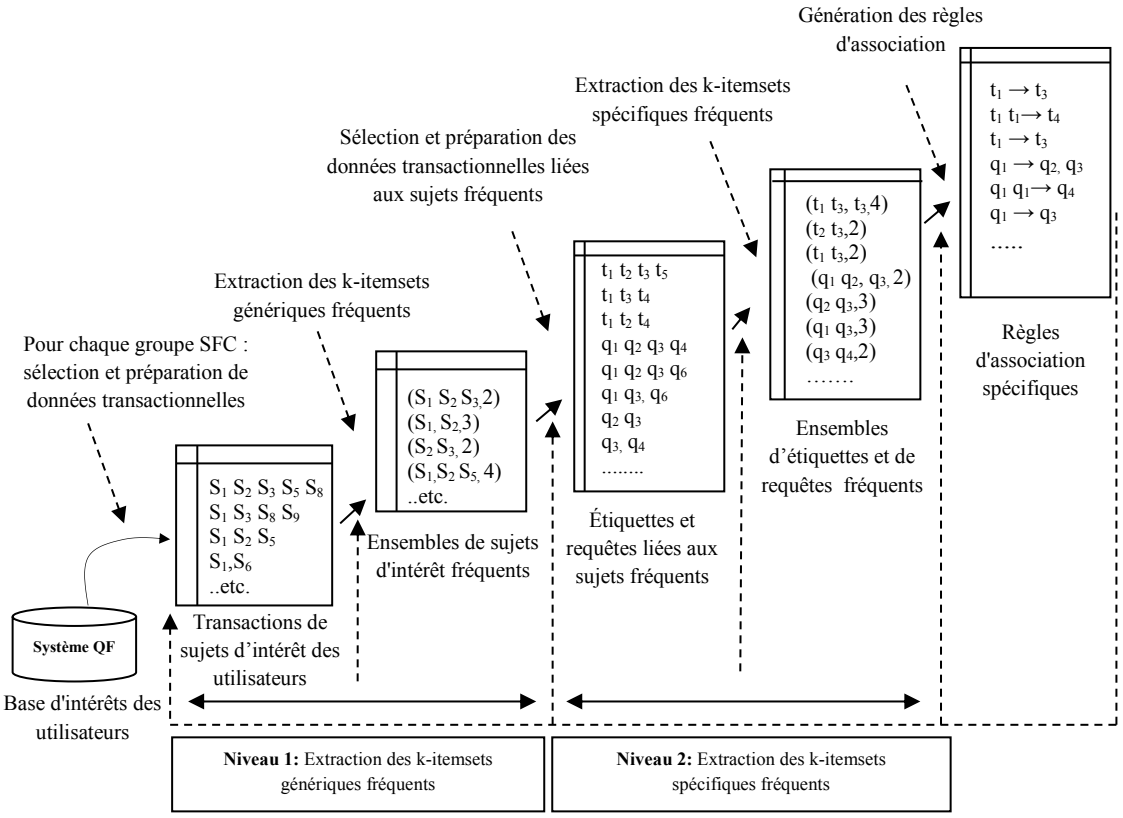


Figure 4. 10. Processus d'extraction des règles d'association à deux niveaux de sélection

Cette technique d'extraction d'itemsets fréquents à deux niveaux de sélection, partant du plus général, les sujets d'intérêt fréquents, au plus spécifique, les objets d'intérêt granulaires fréquents, permet d'enrichir les règles d'association qui sont extraites avec des sujets d'intérêt. Par exemple, durant l'extraction de l'itemset spécifique fréquent (programming, java), le système exploite les sujets d'intérêt à partir desquels ces items spécifiques sont extraits, pour annoter la règle extraite. L'itemset est alors représenté sous la forme suivante : [(programming, suj_i) (java, suj_j)], et la règle extraite est représentée à son tour sous la forme de R1= [(Programming, suj_i) → (Java, suj_j)]. Ces annotations aident à la désambiguïsation des données polysémiques au sein des règles d'association, et aident aussi à faciliter l'accès aux règles qui répondent aux intérêts de chaque utilisateur.

Une fois les règles d'association sont extraites, le système les exploite pour identifier les documents candidats à la recommandation.

Identification des documents candidats à la recommandation : le principe est que le système vérifie pour chaque règle extraite, si les objets antécédents correspondent aux intérêts de l'utilisateur cible ou

sont sémantiquement proches. Si c'est le cas, les documents qui sont liés aux objets conséquents de la règle sont candidats à être recommandés à l'utilisateur.

Afin de faciliter l'exploitation et l'accès aux règles qui répondent aux besoins de chaque utilisateur, nous proposons d'indexer ces règles avec leurs objets antécédents. Une indexation hybride est alors adoptée. Elle permet de décrire les règles avec leur contenu générique et spécifique (cf. tableau 4.4). Par exemple, la règle R1= [(Programming, suj_i) → (Java, suj_j)] est indexée à la fois avec « Programming » et « suj_i ». Cette description facilite l'accès en offrant la possibilité d'effectuer une sélection générique qui sélectionne uniquement les règles qui répondent aux sujets d'intérêt de l'utilisateur puis filtrer celles qui correspondent à ses intérêts spécifiques. Cette indexation peut avoir plusieurs avantages :

- **Améliorer le rappel du système**. Lorsque le système se base sur une sélection générique des règles, la règle d'association « Programmaing→ Java » est aussi considérée comme étant pertinente pour un utilisateur qui n'a pas utilisé l'objet « Programming » dans ses recherches, mais a utilisé d'autres objets qui sont liés au même sujet d'intérêt suj_i. Cela contribue à l'amélioration de la détection de corrélations entre les utilisateurs en les rapprochant à travers les sujets d'intérêt. Nous considérons les objets d'intérêt appartenant à un même sujet de recherche comme étant sémantiquement similaires, le système peut alors étendre la liste des recommandations en proposant aussi à l'utilisateur les documents qui sont liés au même sujet de l'objet « «Java » en conséquence de la règle.
- **Améliorer la précision des résultats par la résolution d'ambigüité lexicale des données polysémiques**. Lorsqu'une telle règle d'association R2 = [(language, suj_i) → (Java, suj_j)] existe, le système l'exploite uniquement pour un utilisateur qui a exploité l'objet « langage » appartenant au sujet suj_i. Et il lui recommande que les documents qui sont liés à l'objet Java qui appartient au sujet suj_j. Cela donne plus de précision aux résultats de recommandation.
- **Améliorer l'efficience système.** L'indexation peut aider à améliorer le temps d'accès aux règles pertinentes de l'utilisateur en éliminant les règles non pertinentes à travers leur description générique, puis filtrer les règles à travers la description spécifique. Cette démarche vise à optimiser le temps d'accès aux règles intéressantes pour cette tache de recommandation.

Donnée spécifique	Règle
programming language	R1 R2

Donnée générique	Règle
suj_1	R1, R2

Tableau 4. 4. Exemple d'indexation hybride des règles d'association

IV.5.4.4.1. Personnalisation du processus d'inférence à base de préférences de l'utilisateur

Afin d'améliorer la prédiction des intérêts de l'utilisateur, le système exploite ses préférences pour sélectionner les règles les plus représentatives de ses intérêts courants. Pour définir ses préférences, trois critères de personnalisation sont exploités comme suit :

1. Sélection du groupe de SFC le plus pertinent pour l'utilisateur. Le système sélectionne parmi les groupes SFC du système, celui qui représente le plus les intérêts courants de l'utilisateur cible, appelé le groupe d'intérêt pertinent. Cette pertinence est multidimensionnelle, elle est évaluée en termes de fraicheur et de fréquence des données. Cette hybridation est fondée sur les motivations suivantes : la fraicheur est prise en considération pour faire face à l'obsolescence des données dans le profil de l'utilisateur. Les besoins de l'utilisateur évoluent au fil du temps et certains intérêts dans son profil ne représenteront plus ses besoins actuels. Dans d'autres cas, les intérêts récents de l'utilisateur peuvent être liés à un besoin spécifique et temporaire qui ne représente pas ses intérêts récurrents. Par exemple, un architecte peut faire des recherches sur le virus de la grippe lorsqu'il a un rhume. Ce besoin est juste temporaire et ne représente pas un centre d'intérêt récurent et pertinent. C'est la raison pour laquelle la fréquence des intérêts est également prise en considération dans cette étape de personnalisation de prédiction et elle est combinée au critère de la fraicheur de données.

Un groupe SFC est dit plus frais pour un utilisateur u_i si le sujet d'intérêt le plus frais dans son profil appartient à ce groupe. Il est considéré comme fréquent si l'utilisateur s'intéresse à plus de sujets appartenant à ce groupe. Cette fréquence est estimée en évaluant le rapport entre le nombre de sujets d'intérêt de cet utilisateur appartenant à ce groupe et le nombre total de ses sujets d'intérêt dans son profil. Ce rapport est évalué à travers la table A du tableau 4.5.

Nous pouvons dire que ce problème de sélection est un problème d'optimisation multi-objectif, et la solution est le groupe SFC_i optimal avec un compromis entre la fraicheur et la fréquence. Nous proposons d'évaluer cette optimalité à travers une fonction pondérée qui combine la fréquence et la fraicheur des données dans le profil de l'utilisateur. Cette fonction linéaire permet d'étudier l'influence de chacun des deux critères sur la prédiction des intérêts des utilisateurs. Ainsi, le degré d'intérêt d'un utilisateur u pour un groupe SFC_i est défini comme suit :

$$score(SFC_i, u) = \alpha * fraicheur\ (u, SFC_i) + (1 - \alpha) fréquence\ (u, SFC_i)$$

Le système sélectionne le groupe SFC_i avec le score maximum. Les règles d'association qui lui sont associées sont exploitées par le système pour l'inférence personnalisée des intérêts de l'utilisateur. Cet ensemble de règles est noté par R_u.

	SFC_1				.	SFC n			
	suj_1	suj_2	.	suj_k	.	suj_1	suj_2	.	suj_m
u_1	1	1	.	1	.	1	0	.	0
u_2	0	1	.		.	1	1	.	1
u_k	0	0	.	1		0	1	.	1

Table A

	SFC_1		..	SFC_n	
	Fraicheur	**Fréquence**	..	**Fraicheur**	**Fréquence**
u_1	X_{11}	Y_{11}	..	X_{1n}	Y_{1n}
u_2	X_{21}	Y_{21}	..	X_{2n}	Y_{2n}
u_k	X_{k1}	Y_{k1}	..	X_{kn}	Y_{kn}

Table B

Pour un utilisateur u_i

	Fraicheur	**Fréquence**
SFC_1	$\boldsymbol{X_{i1}}$	Y_{i1}
..	…	…
SFC_n	$\boldsymbol{X_{im}}$	Y_{i1}

Table C

Tableau 4. 5. Correspondance entre le profil utilisateur et les groupes SFC du système

2. Sélection personnalisée des règles d'association au sein du groupe SFC_i. Il s'agit de sélectionner les règles d'association pertinentes pour l'utilisateur cible au sein de l'ensemble résultant R_u. Cette sélection est fondée aussi sur une pertinence multidimensionnelle qui combine l'aspect fréquentiel et temporel des données. Il s'agit de sélectionner les règles ayant comme antécédents les intérêts optimaux pour l'utilisateur. Cette optimalité est évaluée en combinant les degrés de fréquence et de fraicheur relatifs aux objets antécédents $Ant(R_i)$ de la règle R_i pour sélectionner une ou plusieurs règles les plus pertinentes. Ainsi, la pertinence d'une règle R_i pour un utilisateur u est définie à travers le score $Score(R_i, u)$ comme suit :

$$Score(R_i, u) = score\big(Ant(R_i)\big) = \sum_{j=1}^{N} score\ (obj_j)$$

$$score\ \big(obj_j\big) = \alpha * \ fraicheur\big(obj_j\big) + (1 - \alpha)\ fréquence(obj_j)$$

Où N est le nombre d'objets antécédents dans la règle R_i. Puisque la sélection des règles d'association s'effectue du général, les règles génériques, au spécifique, les règles spécifiques, les objets antécédents d'une règle R_i peuvent représenter des sujets d'intérêt (dans le cas des règles génériques) ou des objets granulaires (étiquettes/requêtes) dans le cas des règles spécifiques. Ainsi, la fréquence et la fraicheur de ces objets sont définies comme suit :

- **Fréquence.** La fréquence d'un sujet d'intérêt dans le profil d'un utilisateur u est évaluée par le degré d'intérêt de cet utilisateur pour ce sujet d'intérêt. Ce degré d'intérêt a été défini dans la section IV.5.4.2. La fréquence d'un objet d'intérêt spécifique obj_i (étiquette/requête) appartenant à un sujet suj_i est mesurée par le nombre d'occurrences tf_{obj_i}de cet objet dans le profil de l'utilisateur u.
- **La fraicheur**. La fraicheur d'un sujet d'intérêt pour un utilisateur u est évaluée à travers l'objet spécifique (document/requête/étiquette) le plus frais appartenant à ce sujet dans le profil de l'utilisateur. La fraicheur d'un objet spécifique a été déjà définie dans la section IV.5.4.2.

La règle ayant le score le plus élevé est exploitée pour prédire les prochains intérêts de l'utilisateur. Les scores obtenus de ces règles peuvent être aussi exploités pour ordonner les règles ce qui permet de proposer des résultats ordonnés.

3. Sélection des documents candidats fondée sur les utilisateurs voisins. Enfin, nous intégrons la similarité entre les utilisateurs pour la sélection personnalisée des documents candidats à la recommandation pour l'utilisateur cible. Il s'agit de sélectionner l'ensemble des utilisateurs ayant exploité les mêmes objets conséquents de la règle résultante, puis évaluer leur similarité avec l'utilisateur cible. Les utilisateurs ayant une grande similarité avec cet utilisateur représentent son voisinage. Ils permettent au système de sélectionner uniquement les documents qui ont été consultés par le voisinage de l'utilisateur cible. Ce processus se résume par les trois étapes suivantes :

- **3.1 Sélection des documents candidats pour la recommandation**. L'ensemble des documents candidats à la recommandation pour un utilisateur cible u est noté par D_u. Cet ensemble englobe les

documents qui sont liés aux objets conséquents de la règle sélectionnée dans l'étape 2, ainsi que les documents connexes par sujets d'intérêt, c'est-à-dire les documents qui sont liés aux mêmes sujets de ces objets conséquents.

- **3.2 Calcul du voisinage par similarité entre utilisateurs**. Le système sélectionne les utilisateurs ayant exploité les mêmes objets conséquents de la règle d'association sélectionnée, puis calcule leurs similarités avec l'utilisateur cible. Cette similarité est évaluée à travers les deux graphes bipartis « utilisateur-document » et « utilisateur-sujet » qui sont représentés sous les deux matrices illustrées dans le tableau 4.6. Ils expriment respectivement les intérêts spécifiques et génériques des utilisateurs.

	suj_1					suj_n			
	d_1	d_2	.	d_k		d_1	d_2		d_j
u_1	V_{11}	v_{12}	.	V_{1k}		V_{11}	v_{12}		V_{1j}
u_2	V_{21}	V_{22}	.	V_{2k}		V_{21}	V22		V_{2j}
u_3	V_{31}	V_{32}	.	V_{3k}		V_{31}	V_{32}		V_{3j}

	d_1	d_2	...	d_k
u_1	v_{11}	v_{12}	..	v_{1k}
u_2	v_{21}	v_{22}	...	v_{2k}
u_3	v_{31}	v_{32}	...	v_{3k}

Table A

	suj_1	...	suj_n
u_1	v_{11}	..	v_{1n}
u_2	v_{21}	...	v_{2n}
u_3	v_{31}	...	v_{3n}

Table B

Tableau 4. 6. Échantillon d'une Q-folksonomie

Deux niveaux de similarité contribuent à l'évaluation de la similarité globale $sim_G\ (u, v)$ entre deux utilisateurs u et v, à savoir i) le niveau générique qui tient compte des sujets de recherche auxquels les utilisateurs s'intéressent, et 2) le niveau spécifique qui considère l'existence et l'absence des documents dans leurs profils. Cette hybridation est considérée pour les motivations expliquées dans la section I.3.2. Pour ce faire, chaque utilisateur u est représenté par deux vecteurs $\overrightarrow{u_1}$ $\overrightarrow{u_2}$ où $\overrightarrow{u_1}$ est le vecteur pondéré qui permet de représenter l'ensemble des documents d'intérêt de l'utilisateur avec leur fraicheur, et $\overrightarrow{u_2}$ est le vecteur pondéré qui représente les degrés d'intérêt de l'utilisateur pour les sujets de recherche (cf. table B du tableau 4.6).

La similarité spécifique sim_1 se base sur la fraîcheur des documents. Cette fraicheur est utilisée dans le but de renforcer la similarité entre deux utilisateurs ayant consommé les mêmes documents dans la même période. La similarité générique sim_2 quant à elle, se base sur la comparaison des degrés d'intérêt

des utilisateurs pour les sujets de recherche. Les deux similarités sont combinées au sein d'une moyenne harmonique qui permet d'évaluer le compromis entre les deux niveaux de similarité considérés. On écrit :

$$sim_G\ (u,v) = \frac{2* sim_1(u,v)*sim_2(u,v)}{sim_1(u,v) + sim_2(u,v)}$$

Ces similarités élémentaires sim_1 et sim_2 sont évaluées chacune en utilisant la mesure du cosinus qui calcule l'angle entre les vecteurs pondérés $\vec{u}.\vec{v}$ qui sont définis respectivement pour les utilisateurs u et v.

$$sim_i(u,v) = Cosinus\ (\overrightarrow{u_i},\overrightarrow{v_i}) = \frac{\overrightarrow{u_i}.\overrightarrow{v_i}}{\|\overrightarrow{u_i}\|^2 * \|\overrightarrow{v_i}\|^2}$$

La mesure de cosinus a prouvé son efficacité avec les vecteurs pondérés au sein de plusieurs travaux. Les auteurs dans (Witten *et al.* 1999) recommandent son utilisation au lieu du produit scalaire. Un seuil de similarité est utilisé pour sélectionner les utilisateurs similaires. Il est défini expérimentalement.

- **3.3 Recommandation des documents candidats**. Les utilisateurs résultants représentent le voisinage de l'utilisateur cible. Il est exploité pour sélectionner les documents intéressants pour l'utilisateur cible. Il s'agit de sélectionner parmi l'ensemble de documents de départ D_u ceux qui ont été consultés par les utilisateurs voisins.

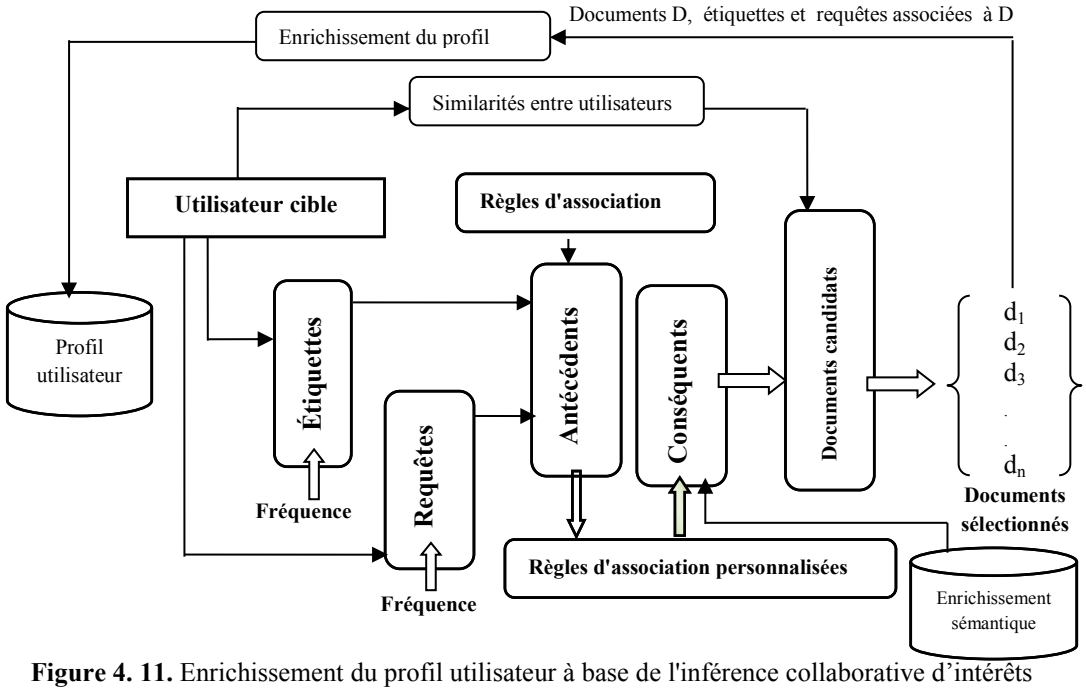


Figure 4. 11. Enrichissement du profil utilisateur à base de l'inférence collaborative d'intérêts

Les documents sélectionnés pour la recommandation contribuent à l'enrichissement automatique du profil de l'utilisateur sans aucun effort de sa part. Dans cette recommandation, le système propose à cet utilisateur de nouvelles expériences de recherche en tenant compte de ses expériences et celles des autres utilisateurs du système. L'utilisateur interagit avec les résultats renvoyés, le processus d'enrichissement du profil à base de ses activités suit son cours à cette étape.

IV.7. Bilan et conclusion

Ce chapitre a présenté un nouveau modèle de représentation d'un profil utilisateur qui prend en compte différents aspects de personnalisation, qui sont susceptibles d'améliorer la pertinence des résultats de recherche. Les contributions de ce modèle couvrent les objectifs que cette étude a fixé pour atténuer les problèmes de la collecte et la représentation des données d'intérêt de l'utilisateur au sein de son profil et la gestion leur évolution au fil du temps. Pour cela, nous avons proposé :

- Une approche topologique qui permet d'analyser les données d'intérêt des utilisateurs dans un système d'extension de folksonomie (Q-folksonomie). Cette analyse tient compte des liens existants dans les graphes bipartis et tripartis qui peuvent être extraits de ce système. Ces graphes permettent d'aider le système à comprendre mieux les intérêts des différents utilisateurs.
- Un modèle multi-niveaux de données qui permet de représenter les intérêts de l'utilisateur sous différents aspects. Pour cela, nous avons fait appel à différents concepts de différents domaines, en l'occurrence i) le domaine du web sémantique pour la conceptualisation du contenu textuel, ii) le domaine d'analyse de corrélation des données pour la délimitation sémantique des intérêts de l'utilisateur en sujets d'intérêt, et iii) du domaine des théories des graphes pour la délimitation contextuelle de ces intérêts. L'avantage de cette représentation est qu'elle est susceptible de contribuer à :
    - La résolution des problèmes d'ambigüités lexicales des données dans le profil de l'utilisateur.
    - L'enrichissement de l'approche topologique par l'ajout de nouvelles dimensions génériques qui contribuent à l'extraction d'éventuelles corrélations (sémantique et contextuelle) entre les données spécifiques des utilisateurs.

- Renforcer les corrélations entre les utilisateurs à travers les hauts niveaux de représentation de leurs intérêts, notamment les sujets d'intérêt et les sujets fortement connexes.

- L'intégration des techniques de feuille de données et de filtrage collaboratif pour l'inférence de nouveaux intérêts pour l'utilisateur. Ces intérêts servent à enrichir davantage son profil. En tirant profit des niveaux de représentation supérieurs des données, le système vise à optimiser le processus d'extraction des données fréquentes dans les profils utilisateurs sur lesquelles se base l'extraction des règles pertinentes pour cette tâche d'inférence.
- L'introduction de l'aspect temporel qui peut être utile pour faciliter le traitement des données dans le profil de l'utilisateur. Cela est fait en définissant les préférences de l'utilisateur en termes de fraicheur de données. Cet aspect temporel peut être aussi utile pour renforcer la similarité entre les utilisateurs qui ont des comportements similaires dans des périodes de temps proches.

La majeure limitation de cette approche d'inférence d'intérêts est le démarrage à froid d'un nouvel utilisateur. Ce problème se traduit par le fait qu'un système ne peut recommander des données aux utilisateurs avec des profils vides. Le chapitre 6 discute une approche qui vise à atténuer ce problème.

Le chapitre prochain présente des modèles de recherche d'information qui intègrent le profil de l'utilisateur conçu pour personnaliser ses recherches. Cette personnalisation vise à améliorer la pertinence des résultats.

Chapitre 5 : Approche hybride de personnalisation de recherche d'information

V. 1. Introduction

La personnalisation de données orientée-utilisateur consiste à intégrer le profil de cet utilisateur dans le SRI afin de l'aider à accéder aux informations qui sont liées à ses préférences et à ses intérêts, et/ou à ceux des autres utilisateurs similaires. Le chapitre précédent a montré comment ces données d'intérêt sont recueillies, représentées, et enrichies au sein d'un modèle générique de profil utilisateur. Dans ce chapitre, nous proposons deux méthodes de personnalisation qui les exploitent en deux façons différentes pour répondre aux besoins spécifiques de l'utilisateur exprimés par des requêtes de recherche. Il consiste à définir pour chacune de ces deux méthodes proposées, le niveau et la technique d'intégration de ces intérêts dans le système de recherche en vue de répondre à la pertinence des résultats (cf. section I.1.4). Une analyse comparative entre nos propositions et quelques principaux travaux de la littérature est ensuite présentée. Elle est effectuée en termes de plusieurs critères, tels que les types d'intérêts qui sont exploités dans cette personnalisation, les ressources exploitées (couche de traitements et espace mémoire requis pour cette personnalisation), et les stratégies de recherche qui sont offertes par ces différentes techniques.

V. 2. Intégration du profil utilisateur

Nous avons vu dans les précédents chapitres (chapitre 2 et 3) qu'un SRI se compose de trois principales tâches, à savoir, l'indexation des documents, l'interprétation de la requête utilisateur et le processus de mise en correspondance « requête-document ». Ce processus localise le contenu de la requête dans l'index documentaire et retourne les documents triés selon leur score de correspondance. La question importante est de savoir comment les données d'intérêt de l'utilisateur peuvent être intégrées dans le système pour mieux influencer et accélérer la RI. Selon la littérature, ces intérêts peuvent être intégrés selon diverses façons et à différentes étapes du processus RI. En effet, l'intégration peut être effectuée au niveau de la représentation des documents (Bouadjenek *et al.* 2013) (Bouhini *et al.* 2013b), durant le traitement de la requête (De Meo *et al.* 2010; Bouhini *et al.* 2016) (Zhou *et al.* 2017), ou dans le réordonnancement des documents résultats (Daoud *et al.* 2010b) (cf. section II.2.2). Dans ce chapitre,

nous proposons deux techniques d'intégration de ces intérêts dans le SRI. La première les intègre au niveau de la description des documents dans l'index, et la deuxième les exploite dans le réordonnancement des résultats. Ces techniques tiennent compte en plus des intérêts individuels de l'utilisateur, de ceux des autres utilisateurs similaires. Pour cela, une technique de détection des utilisateurs similaires est également proposée. Ces utilisateurs définissent le voisinage de l'utilisateur cible, ils sont appelés aussi les utilisateurs voisins. Ce voisinage évolue au fur et à mesure de l'évolution des intérêts des utilisateurs. Une technique de détection automatique de ce voisinage est donc proposée.

V.2.1. Démarche I: description personnalisée des documents

Tel qu'il a été vu dans le chapitre précédent, les requêtes de recherche et les étiquettes d'annotation décrivent les besoins de l'utilisateur dans son profil. Ces deux entités décrivent également les documents dans le système d'analyse de données QF (cf. figure 4.3) (Hannech *et al.* 2016c). En considérant cela, ces données d'intérêt sont exploitées dans notre SRI pour l'enrichissement de l'index documentaire. Dans le chapitre 3, nous avons proposé un modèle d'indexation multidimensionnelle qui permet de décrire les documents selon plusieurs espaces de données. Parmi les espaces proposés, nous citons l'espace social qui décrit le contenu de ces documents avec les étiquettes d'annotation qui sont employées par différents utilisateurs. Ces étiquettes sont toutes mélangées dans l'index sans faire la distinction entre celles qui proviennent d'un même utilisateur et sans faire référence à ces utilisateurs (cf. figure 5.1 partie 1). Cela ne permet d'avoir qu'une seule description commune pour tous les utilisateurs.

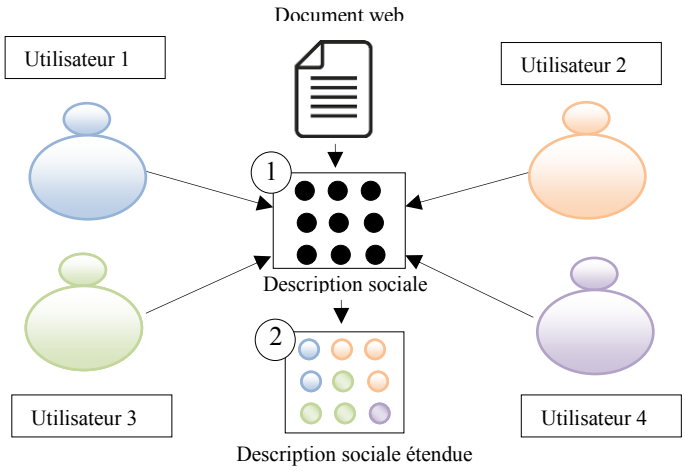


Figure 5. 1. Extension de la description sociale d'un document

Dans ce chapitre, nous proposons une extension de ce modèle d'indexation qui permet d'apporter un aspect descriptif centré utilisateur à l'univers de représentation des documents (cf. figure 5.1 partie 2). Il consiste à décrire ces documents selon chaque utilisateur, ce qui aide à filtrer leurs résultats de recherche selon leurs intérêts. Cela est possible en enrichissant l'univers d'indexation avec les requêtes de recherche et les étiquettes de chaque utilisateur. Ceci aide à promouvoir les documents dont le contenu correspond aux intérêts récurrents de l'utilisateur.

Pour illustrer l'idée générale de cette contribution, considérons l'exemple suivant. Supposons que deux utilisateurs u_1 et u_2 souhaitent chacun obtenir des informations à propos des logiciels et des conceptions des logiciels. Ils soumettent alors la requête « software design ». L'utilisateur u_1 est un informaticien qui s'intéresse aux processus de conception des logiciels. L'utilisateur u_2 quant à lui est un architecte, il est donc beaucoup plus intéressé par les logiciels destinés à la conception des maquettes, des bâtiments, etc. Comme nous pouvons le voir, malgré le fait que ces utilisateurs ont formulé la même requête, leur besoin d'information est différent. Nous supposons que nous avons des documents d_1, d_2 et d_3 qui contiennent un nombre d'occurrences différent des jetons de la requête (cf. tableau 5.1)

Entité/jeton	Software	Design
Requête	1	1
d_1	26	4
d_2	29	3
d_3	8	72
u_1	15	5
u_2	3	12

Tableau 5. 1. Distribution des jetons dans la requête, les documents et les profils des utilisateurs

En se basant sur la distribution des jetons du tableau 5.1, un SRI classique renverrait pour les deux utilisateurs la même liste de documents qui répond à leur requête de recherche et dans le même ordre basé sur leur score de pertinence. Tandis qu'un SRIP, qui prend en compte les préférences des utilisateurs stockées dans leurs profils, considère que le jeton « Software » de la requête est plus important pour u_1 par rapport au jeton « Design » et inversement pour u_2. Ainsi, il considère que les documents d_2 et d_1 sont plus pertinents que d_3 pour l'utilisateur u_1, et d_3 est plus pertinent que d_1 et d_2 pour l'utilisateur u_2.

Les questions qui pourraient être posées maintenant sont les suivantes : comment formaliser une représentation personnelle d'un document dans un cadre hybride[8], multidimensionnel[9] et collaboratif[10], qui aide à l'adaptation de l'index documentaire aux intérêts des utilisateurs et contribue à personnaliser leur recherche ? Quelle stratégie adaptative de l'index doit être définie afin de tenir compte de ce formalisme de représentation dans la description globale des documents ? Cette formalisation nécessite en premier lieu, la préparation des données d'intérêt pertinentes pour cette adaptation, et en second lieu la proposition d'une technique de combinaison qui tient compte de la nouvelle représentation.

V.2.1.1. Préparation des données pour une représentation personnalisée du document

Cette étape consiste à déterminer les données pertinentes pour enrichir la description des documents dans l'index multidimensionnel. Cette tâche est effectuée en plusieurs étapes :

1. Préparation des données d'intérêt de chaque utilisateur qui vont être exploitées dans cette tâche d'enrichissement. Ainsi, deux ensembles de données sont considérés pour chaque utilisateur u_j, à savoir, l'ensemble de ses requêtes de recherche noté par Q_{u_j} qui est exploité dans l'enrichissement des documents au sein des deux espaces, identitaire et sémantique, de l'index, et l'ensemble de ses étiquettes d'annotation noté par T_{u_j} qui est exploité au sein de l'espace social.

$$Q_{u_j} = \left\{ q_m \in Q \text{ pour m} \in [1, |Q_{u_j}|] \,/\, \exists\, d_n \in \mathrm{D} : < \mathrm{u_j}, q_m, d_n > \in R_1, R_1 \subseteq \mathrm{U} \times \mathrm{Q} \times \mathrm{D} \right\}$$

$$T_{u_j} = \left\{ t_i \in T \text{ pour i} \in [1, |T_{u_j}| \,/\, \exists\, d_n \in \mathrm{D} : < u_j, t_i, d_n > \in R_2, R_2 \subseteq \mathrm{U} \times \mathrm{T} \times \mathrm{D} \right\}$$

Où R_1 et R_2 sont les relations qui relient les entités U, T, Q, D dans le système QF.

2. Sélection des données d'enrichissement de chaque utilisateur u_j pour chaque document d_k. Ce sont les objets de contenu (requêtes et étiquettes) appartenant aux deux ensembles de données Q_{u_j} et T_{u_j} et correspondant au contenu du document cible d_k dans l'index multidimensionnel. Chaque objet est noté par « $obj_{d_k}^{u_j}$ ». Il est à noter que quand on fait référence au contenu d'un document cela ne désigne pas

[8] Hybride : où le document peut être lié à plusieurs entités informationnelles : facette de contenu (espace), requête, étiquette.
[9] Multidimensionnel : où l'univers de représentation est multi-espaces.
[10] Collaboratif : où le document peut être représenté selon plusieurs utilisateurs.

uniquement son contenu textuel, mais tout contenu le décrivant dans l'index multidimensionnel (cf. section III.3.1.1). Ainsi, trois ensembles de données sont considérés :

- **Données d'adaptation du document d_k au sein de l'espace identitaire, elles sont notées par $IES_{d_k}^{u_j}$ (pour Identity Extension Set)**: nous avons vu dans la section III.3 que le contenu d'un document ainsi que celui d'une requête de recherche sont représentés par un ensemble de jetons. Cette étape d'enrichissement consiste donc à extraire les jetons qui sont à la fois dans l'ensemble Q_{u_j} de l'utilisateur u_j et dans le contenu identitaire du document d_k noté par IS_{d_k}.

$$IES_{d_k}^{u_j} = \left\{ tk \in \left(IS_{d_k} \cap Q_{u_j} \right) \right\}$$

- **Données d'adaptation du document d_k au sein de l'espace social, elles sont notées par $SES_{d_k}^{u_j}$ (pour Social Extension Set)**: contrairement aux requêtes de recherche, la longueur d'une étiquette ne dépasse généralement pas deux mots et elle est le plus souvent de cardinalité égale à 1. Une étiquette t_i n'est donc pas tokenisée dans notre modèle. Ainsi, les données d'adaptation $SES_{d_k}^{u_j}$ d'un utilisateur u_j pour un document d_k sont les étiquettes qui sont à la fois dans l'ensemble d'annotations T_{d_k} associé au document d_k et l'ensemble des annotations T_{u_j} qui ont été employées par cet utilisateur pour annoter n'importe quel document.

$$SES_{d_k}^{u_j} = \left\{ t \in (T_{d_k} \cap T_{u_j}) \right\}$$

Où T_{d_k} est l'ensemble de toutes les étiquettes qui sont associées dans le système QF au document d_k par tous les utilisateurs U. On écrit :

$$T_{d_k} = \bigcup_{i=1, u_i \in U}^{|U|} T_{d_k}^{u_i}$$

Où $T_{d_k}^{u_j}$ est l'ensemble des annotations qui sont associées à un document d_k par un utilisateur u_i.

- **Données d'adaptation du document d_k au sein de l'espace sémantique, elles sont notées par $SeS_{d_k}^{u_j}$ (pour Semantic Extension Set)**: ce sont les jetons qui sont à la fois dans l'ensemble Q_{u_j} et dans l'ensemble des jetons sémantiques Se_{d_k} qui décrivent le document d_k dans l'espace sémantique.

$$SeS_{d_k}^{u_j} = \left\{ tk \in \left(Se_k \cap Q_{u_j} \right) \right\}$$

Les deux ensembles Q_{u_j} et T_{u_j} peuvent contenir plusieurs occurrences d'un même objet (requête ou étiquette). Cela permet de calculer la fréquence d'occurrence de chaque objet dans les trois ensembles d'enrichissement obtenus $IES_{d_k}^{u_j}, SES_{d_k}^{u_j}$ *et* $SeS_{d_k}^{u_j}$.

3. La sélection des objets d'enrichissement pertinents qui reflètent réellement le contenu des documents. Ainsi, deux paramètres de pertinence sont considérés : le contexte de l'objet et sa représentativité pour le contenu du document cible.

- **Contexte de l'objet :** un objet d'enrichissement $obj_{d_k}^{u_j}$ est pris en compte dans la description d'un document d_k que s'il appartient au même cluster contextuel cd_i de ce document qui leur est associé dans le système d'analyse QF. Cela permet de ne pas considérer dans la description du document l'occurrence des objets ayant différentes interprétations et ne correspondant pas à son contexte de recherche. Par exemple, lorsque le jeton « Virus » apparait dans l'historique des requêtes ou d'annotations de l'utilisateur, il ne doit pas être utilisé pour enrichir tous les documents qui sont décrits avec ce jeton, car ils peuvent couvrir différentes interprétations (cf. section I.1.4). On écrit :

$$IES_{d_k}^{u_j} = \left\{ tk^{q_m} \in \left(IS_{d_k} \cap Q_{u_j} \right) / \exists cd \in CD : d_k R_3 cd \; et \; q_m R_4 cd \right\}$$

$$SES_{d_k}^{u_j} = \left\{ t \in \left(T_{d_k} \cap T_{u_j} \right) / \exists cd \in CD : d_k R_3 cd \; et \; t \; R_5 cd \right\}$$

$$SeS_{d_k}^{u_j} = \left\{ tk^{q_m} \in \left(Se_{d_k} \cap Q_{u_j} \right) / \exists cd \in CD : d_k R_3 cd \; et \; q_m R_4 cd \right\}$$

Où ${tk_i}^{q_m}$ fait référence à un jeton tk_i qui a été extrait d'une requête q_m. CD est l'ensemble de tous les clusters contextuels dans le système QF.

Si aucune relation ne relie le document cible d_k à une classe contextuelle dans le système d'analyse QF, cela indique que ce document n'a toujours pas été considéré comme pertinent par un utilisateur donné. Afin d'attribuer une classe contextuelle à ce document qui aide à sélectionner ses objets d'enrichissement, le système sélectionne un document d_j appartenant au système QF qui soit en relation sémantique R'avec le contenu du document cible d_k. Cette relation sémantique peut être évaluée à travers le calcule de cosinus de leurs vecteurs respectifs extraits de leurs contenus identitaires. On écrit :

$d_k\ R' d_j$ si et seulement si cosinus $(\overrightarrow{d_k}, \overrightarrow{d_j}) \geq \Omega$ où Ω est un seuil minimal de similarité

- **Représentativité de l'objet, elle est notée par $\boldsymbol{Rep_{obj_i}^{d_k}}$**: un objet d'enrichissement $obj\ _{d_k}^{u_j}$ est retenu pour l'enrichissement d'un document d_k que s'il est descriptif de son contenu dans l'espace d'enrichissement cible, notamment, dans l'espace identitaire ou sémantique pour le cas des requêtes d'enrichissement, et dans l'espace social pour le cas des étiquettes. Deux cas de figure se présentent :
  - Si l'objet d'enrichissement est lié à la fois au document cible et à l'utilisateur cible u_j dans le système QF, il est considéré comme représentatif de son contenu, car la pertinence de ce document pour cet objet a déjà été considérée par u_j du système lors de ses activités de recherche. Cette représentativité est donc basée sur la pertinence de l'utilisateur. On écrit :

    $obj\ _{d_k}^{u_j}$ est dit représentatif d_k si et seulement si $<u_j, obj\ _{d_k}^{u_j}, d_k> \in R_1$ ou $<u_j, obj\ _{d_k}^{u_j}, d_k> \in R_2$ tel que $R_1 \subseteq \mathrm{U} \times \mathrm{Q} \times \mathrm{D}$ et $R_2 \subseteq \mathrm{U} \times \mathrm{T} \times \mathrm{D}$.
  - Si l'objet n'est pas lié au document cible d_k dans le système QF, sa représentativité pour le contenu de ce document sera évaluée à base de la pertinence système. Cette pertinence se résume par l'idée suivante : l'objet est utilisé pour interroger notre SRI multi-facettes (SRIF), si le document cible d_k appartient au top k documents retournés par le système derrière cette recherche, l'objet est considéré comme pertinent pour le document et aussi représentative de son contenu. Dans le cas contraire, l'objet est considéré comme non pertinent et ne sera donc pas retenu (cf. Annexe 2). On écrit :

$$IES_{d_k}^{u_j} = \left\{tk \in IES_{d_k}^{u_j} / \ \ Rep_{tk_i}^{d_k} = vrai\right\}, SES_{d_k}^{u_j} = \left\{t \in SES_{d_k}^{u_j} / \ \ Rep_{t_i}^{d_k} = vrai\right\}$$

$$SeS_{d_k}^{u_j} = \left\{tk \in SeS_{d_k}^{u_j} / \ \ Rep_{tk_i}^{d_k} = vrai\right\}$$

V.2.1.2. Adaptation orientée utilisateur de l'index documentaire

Cette étape propose une méthode qui intègre dans l'index multidimensionnel la représentation personnalisée des documents. Dans notre modèle d'indexation, les documents sont organisés en plusieurs espaces et sont structurés au sein de chaque espace en champs de description. Cela permet au système

d'effectuer des recherches personnalisées qui ciblent un ou plusieurs espaces d'indexation, cibler ou un ou plusieurs champs de contenu au sein d'un espace donné. Il permet également de privilégier certains d'entre eux selon le besoin d'un ou plusieurs utilisateurs ou selon un paramètre de personnalisation défini par le système. Cela est effectué en appliquant des coefficients de pondération. Ainsi, pour cette adaptation orientée utilisateur, chaque document est enrichi avec plusieurs champs où chacun référence les intérêts d'un utilisateur (ses requêtes de recherche au sein des deux espaces, identitaire et sémantique, et ses étiquettes au sein de l'espace social). Il s'agit d'étendre la description de chaque document d_k au sein de son univers de description avec les intérêts de chaque utilisateur. La figure 5.2 illustre une instance de cette représentation étendue pour un document d_k.

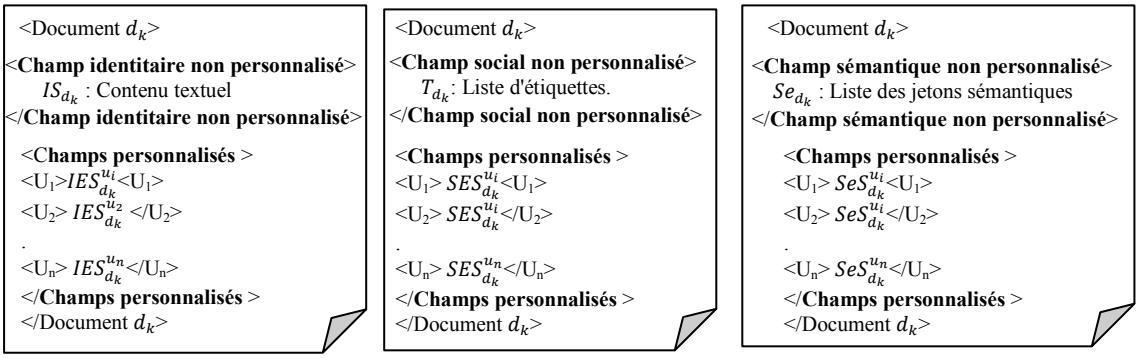


Figure 5. 2. Représentation étendue d'un document dans un index multidimensionnel

Ainsi, l'univers de représentation $\overrightarrow{\mathrm{U}}_{\mathrm{d_k}}$d'un document $\mathrm{d_k}$ peut être formalisé en fonction de ses trois différents espaces de description. Chaque espace est constitué d'un ensemble de jetons qui décrivent son contenu selon un aspect différent: identitaire, social et sémantique (cf. section III.3). On écrit :

$$\overrightarrow{\mathrm{U}}_{\mathrm{d_k}} = (\overrightarrow{\mathrm{S1}}_{\mathrm{d_k}}, \overrightarrow{\mathrm{S2}}_{\mathrm{d_k}}, \overrightarrow{\mathrm{S3}}_{\mathrm{d_k}}), \text{ où}$$

- $\overrightarrow{\mathrm{S1}}_{\mathrm{d_k}}$ est la description identitaire étendue. Elle est constituée de deux dimensions : la dimension identitaire personnalisée $\overrightarrow{IES}_{\mathrm{d_k}}$ et la dimension identitaire non personnalisée $\overrightarrow{IS}_{\mathrm{d_k}}$, tel que :

$$\overrightarrow{\mathrm{S1}}_{\mathrm{d_k}} = (\overrightarrow{\mathrm{IES}}_{\mathrm{d_k}} \cup \overrightarrow{\mathrm{IS}}_{\mathrm{d_k}}) = (\coprod_{j=1}^{|U|} IES_{d_k}^{u_j} \cup \coprod_{i=1,tk_i \in IS_{d_k}}^{|IS_{d_k}|} tk_i)$$

- $\overrightarrow{\mathrm{S2}}_{\mathrm{d_k}}$ est la description sociale étendue du document. Elle est constituée de la dimension sociale personnalisée $\overrightarrow{SES}_{\mathrm{d_k}}$ et de la dimension sociale non personnalisée $\overrightarrow{SS}_{\mathrm{d_k}}$, tel que :

$$\overrightarrow{\mathrm{S2}} = (\overrightarrow{\mathrm{SES}}_{\mathrm{d_k}} \cup \overrightarrow{\mathrm{SS}}_{\mathrm{d_k}}) = (\coprod_{j=1}^{|U|} SES_{d_k}^{u_j} \cup T_{d_k})$$

- $\overrightarrow{\mathrm{S3}}_{\mathrm{d_k}}$ est la description sémantique du document. Elle est constituée de son côté aussi de la dimension sémantique personnalisée $\overrightarrow{SeS}_{\mathrm{d_k}}$ et de la dimension sémantique non personnalisée $\overrightarrow{Se}_{\mathrm{d_k}}$ qui englobe les jetons qui sont liés sémantiquement au contenu identitaire $\overrightarrow{IS}_{\mathrm{d_k}}$ ou au contenu social $\overrightarrow{SS}_{\mathrm{d_k}}$.

Dans les sections prochaines, nous allons voir plus de détails sur la représentation des documents dans l'index multidimensionnel étendu, notamment le principe de pondération des jetons dans l'univers d'indexation et le modèle adopté pour cette pondération.

V.2.1.2.1. Principe de la pondération des jetons d'indexation

Un document est indexé en attribuant des scores de pondération aux jetons qui décrivent son contenu. Ces pondérations représentent leur importance relativement à ce contenu. Elles se basent généralement sur le calcul des occurrences de ces jetons et sur d'autres propriétés qui aident à calculer le score global d'un document par rapport à une requête de recherche. Il existe plusieurs modèles de pondération dans la littérature (cf. section II.1.2.2). Nous avons choisi le modèle BM25F, il est souvent utilisé dans le cas des documents qui sont structurés en plusieurs champs d'information où « F » fait référence aux différents champs (Fields) qui constituent le document (Lu *et al.* 2006) (Pérez-Agüera *et al.* 2010). Ce modèle répond exactement à la structure de nos documents dans l'univers d'indexation. Nous commençons dans cette section par donner un rappel sur la fonction de pondération W_{d_k,tk_i} d'un jeton tk_idans un document d_k structuré en plusieurs champs F. Nous présentons dans la section qui suit le nouveau modèle de pondération étendue.

$$W_{d_k,tk_i} = TF_{d_k,tk_i} \times IDF_{tk_i} \quad (5.\ 1)$$

Tel que :

$$TF_{d_k,tk_i} = \frac{TF_{Fd_k,tk_i}}{k1 + TF_{Fd_k,tk_i}} \quad (5.\ 2)$$

$$TF_{Fd_k,tk_i} = \sum_{f \in d_k} \alpha_f \times \overline{tf_{fd_k,tk_i}} \quad (5.\ 3)$$

Avec:

- TF_{Fd_k,tk_i} est la fréquence d'occurrence du jeton tk_i dans l'ensemble des champs F du document d_k

- IDF_{tk_i}représente l'importance du jeton dans le corpus documentaire.
- k_1 représente le paramètre classique dans le modèle BM25. Il permet de contrôler le taux de saturation de la fréquence des jetons.
- α_f: représente le poids attribué au champ f du document d_k.
- tf_{fd_k,tk_i}est la fréquence d'occurrence du jeton tk_i dans le champ f du document d_k.
- $\overline{\qquad tf_{fd_k,tk_\iota}}$ est la version normalisée de la fréquence tf_{fd_k,tk_i}.

$$\overline{tf_{fd_k,tk_\iota}} = \frac{tf_{fd_k,tk_i}}{(1-b_f) + b_f \frac{fl}{avg(fl)}} \qquad (5.4)$$

Avec :

- b_f: correspond au paramètre b dans le modèle classique BM25 pour le champ f du document d_k. Il permet de contrôler la normalisation en fonction de la taille du champ f.
- fl et $avgfl$ sont respectivement la taille du champ et la taille moyenne des champs du document d_k.

V.2.1.2.2. Un nouveau modèle de pondération étendue des jetons d'indexation

Dans notre cas, un document d_k est décrit dans l'index multidimensionnel en fonction de plusieurs espaces. Chaque espace $\overrightarrow{\mathrm{S}_{\iota \mathrm{d_k}}}$ est constitué d'un ensemble de champs ($f_1^{S_i}$,.., $f_m^{S_i}$). Chaque champ $f_j^{S_i}$ de cet espace est représenté par un ensemble de jetons pondérés. Ces pondérations décrivent leur importance dans ledit champ décrivant ce document. On écrit :

$$\vec{\mathrm{U}}_{\mathrm{d_k}} = (\overrightarrow{\mathrm{S}_{1\mathrm{d_k}}}, \overrightarrow{\mathrm{S}_{2\mathrm{d_k}}}, \overrightarrow{\mathrm{S}_{3\mathrm{d_k}}}), \text{ où}$$

$$\overrightarrow{\mathrm{S}_{\iota \mathrm{d_k}}} = \left\{ \overbrace{((tk_1^{f_1^{S_i}}, W_{d_k,tk_1}^{f_1^{S_i}}), .. (tk_i^{f_1^{S_i}}, W_{d_k,tk_i}^{f_1^{S_i}}))}^{f_1^{S_i}}, \ldots, \overbrace{((tk_1^{f_m^{S_i}}, W_{d_k,tk_1}^{f_m^{S_i}}), .. (tk_j^{f_m^{S_i}}, W_{d_k,tk_j}^{f_m^{S_i}}))}^{f_m^{S_i}} \right\}$$

L'objectif derrière cette démarche d'enrichissement qui consiste à intégrer au sein de la description globale du document d_k des champs de description relatives à chaque utilisateur, est d'augmenter la fréquence d'occurrence des jetons TF_{d_k,tk_i} qui correspondent aux d'intérêt des utilisateurs. Il s'agit d'ajuster les scores de pondération des jetons dans l'index documentaire quand ils correspondent aux

intérêts de l'utilisateur. Ceci permet d'impacter le score de pertinence $RSV\ (q, d_k, u)$ d'un document d_k pour une requête de recherche q avec le contenu étendu qui correspond à un ou plusieurs utilisateurs. La pondération obtenue d'un jeton tk_i dans un document d_k est appelée dans notre modèle par la pondération étendue. Elle est définie en fonction d'un espace S_i et des champs de description qui décrivent le document dans cet espace. Elle est notée par $EW_{d_k,tk_i}^{S_i}$ et est formalisée comme suit :

$$EW_{d_k,tk_i}^{S_i} = ETF_{d_k,tk_i}^{S_i} \times EIDF_{tk_i}^{S_i} \quad (5.\ 5)$$

Tel que :

$$ETF_{d_k,tk_i}^{S_i} = \frac{ETF_{Fd_k,tk_i}^{S_i}}{k1 + ETF_{Fd_k,tk_i}^{S_i}} \quad (5.\ 6)$$

$$ETF_{Fd_k,tk_i}^{S_i} = \sum_{(f^{Si}) \in S_{i_{dk}}} \alpha_{f^{Si}} \times \overline{tf_{f_{d_k,tk_i}^{Si}}} \quad (5.\ 7)$$

Avec :

- $ETF_{Fd_k,tk_i}^{S_i}$ est la fréquence d'occurrence étendue du jeton tk_i dans l'ensemble des champs F du document d_k qui le décrivent dans l'espace S_i.
- $EIDF_{tk_i}^{S_i}$ représente l'importance de tk_i dans l'ensemble des documents appartenant à l'espace S_i.
- $tf_{f_{d_k,tk_i}^{Si}}$ est la fréquence d'occurrence de tk_i dans un champ f^{Si} qui décrit le document d_k dans l'espace S_i.

Les champs qui sont considérés dans le calcul de cette fréquence étendue $ETF_{Fd_k,tk_i}^{S_i}$ au sein de l'espace identitaire S_1 sont les suivants:

- Le champ « contenu textuel » du document noté par f_c^{S1}. Il représente son contenu identitaire IS_{d_k}.
- Les champs de description du document d_k par rapport aux profils des utilisateurs. Ils sont représentés par l'union des différents champs qui représente chacun les intérêts d'un utilisateur u_j $(\bigcup_{j=1}^{|U|} f_{u_j,d_k,tk_i}^{S1})$. Ainsi la fréquence d'occurrence étendue ETF_{Fd_k,tk_i}^{S1} est la fréquence d'occurrence globale qui tient compte de tous ces champs de description. On écrit :

$$ETF_{Fd_k,tk_i}^{S1} = \alpha_c \times \overline{tf_{f_c^{S1}d_k,tk_i}} + \sum_{j=1}^{|U|} \alpha_{u_j} \times \overline{tf_{f_{u_j,d_k,tk_i}^{S1}}} \quad (5.\ 8)$$

- α_c, α_{u_j} représentent respectivement les poids attribués au champ identitaire et ceux des utilisateurs système qui décrivent d_k dans S1.

L'évaluation de l'occurrence d'un jeton tk_i au sein de l'espace social S_2 quant à elle implique la considération des champs suivants:

- Le champ social f_s^{S2} qui représente l'ensemble de toutes les annotations T_{d_k} associé au document.
- Les champs de la description étendue du document par rapport aux profils des utilisateurs. On écrit:

$$ETF_{Fd_k,t_i}^{S2} = \alpha_s \times \overline{tf_{f_c^{S2}d_k,t_\iota}} + \sum_{j=1}^{|U|} \alpha_{u_j} \times \overline{tf_{f_{u_j}^{S2}d_k,t_\iota}} \quad (5.9)$$

De la même façon, l'occurrence des jetons est évaluée au sein de l'espace sémantique S_3 en fonction des champs de description qui sont définis dans la représentation multidimensionnelle du document (cf. figure 5.2).

De cette façon, les jetons qui sont souvent employés par l'utilisateur pour exprimer ses besoins via des requêtes de recherche ou des étiquettes d'annotation obtiennent une pondération étendue élevée et permettent de promouvoir les documents auxquels ils appartiennent. On écrit :

$$EW_{d_k,tk_i}^{S_i} = \overbrace{\frac{gtf_{d_k,tk_i}^{S_i}}{k1 + gtf_{d_k,tk_i}^{S_i}}}^{ETF_{d_k,tk_i}^{S_i}} * \log \overbrace{\left(\frac{N - df_{tk_i}^{S_i} + \mathrm{b}}{df_{tk_i}^{Si} + \mathrm{b}}\right)}^{EIDF_{tk_i}^{S_i}} \quad (5.10)$$

Où $df_{tk_i}^{Si}$ est le nombre de documents dans l'espace S_i qui contiennent le jeton tk_i. $k1$ et b sont des constantes utilisées dans le modèle classique BM25F. Elles sont souvent fixées respectivement à 1.2 et 0.5.

V.2.1.2.3. Modèle de pondération centrée utilisateur

Ce modèle de pondération permet de définir le poids personnalisé $PW_{d_k,tk_i,u_j}^{S_i}$d'un jeton tk_i dans un document d_k selon un utilisateur donné u_j. Il consiste à cibler les champs qui décrivent ce document selon cet utilisateur u_j et son voisinage noté par RS_{u_j}. On écrit:

$$PW_{d_k,tk_i,u_j}^{S_i} = PTF_{d_k,tk_i,u_j}^{S_i} \times PIDF_{tk_i}^{S_i} \quad (5.11)$$

$$PTF_{d_k,tk_i,u_j}^{S_i} = \alpha_{f^{S_i}} \times \overline{tf_{f_{d_k,tk_\iota}^{S_\iota}}} + \alpha_{u_j} \overline{tf_{f_{d_k,tk_\iota,u_j}^{S_\iota}}} + \alpha_{RS_{u_j}} \sum_{u_n \in RS_{u_j}, n=1}^{|RS_{u_j}|} \overline{tf_{f_{d_k,tk_\iota,u_n}^{S2}}} \quad (5.12)$$

- f^{S_i} est le champ de description non personnalisé d'un document au sein de l'espace S_i. Il fait référence au champ identitaire f_c ou social f_s ou sémantique f_{se} selon l'espace dans lequel cette pondération est évaluée.
- f_{u_j} est le champ de description du document d_k relatif à l'utilisateur cible.
- f_{d_k,tk_i,u_n}^{Si} est les champs de description du document d_k relatifs à un utilisateur u_n appartenant au voisinage de l'utilisateur cible. Dans la section prochaine, nous définissons le modèle de construction de ce voisinage ainsi que nos motivations pour cette exploitation.

V.2.1.2.4. Modèle de construction du voisinage utilisateur

A. Définition et motivations

La notion du voisinage est utilisée pour indiquer un groupe d'utilisateurs qui sont similaires à un utilisateur donné. Ce voisinage est intégré dans notre étude pour la définition d'une représentation personnalisée des documents selon un utilisateur donné. Cette représentation du document est considérée comme étant dynamique puisque le contenu du voisinage intégré est de son côté dynamique. Elle est utilisée à la volée lors de l'interrogation de l'index documentaire en personnalisant l'accès sur un sous-ensemble de champs qui décrivent les documents selon l'utilisateur cible et son voisinage. L'intégration de ce voisinage est basée sur les motivations suivantes: considérer seulement les intérêts individuels de l'utilisateur dans la représentation des documents peut engendrer une baisse du rappel système lorsque la recherche actuelle de cet utilisateur représente un nouveau besoin en information pour lui ou elle est exprimée différemment par rapport à ses précédents intérêts. Le système peut alors ignorer les documents qui couvrent ce besoin et qui ne correspondent pas aux intérêts précédents de cet utilisateur. Ces documents peuvent être intéressants pour cet utilisateur. Dans de tels cas, le voisinage de l'utilisateur peut aider à enrichir cette représentation en bénéficiant des intérêts des autres utilisateurs similaires pour étendre la représentation personnalisée des documents. Aussi, un document peut avoir un score de correspondance faible lorsqu'il est mal décrit par l'utilisateur cible (le cas de la description sociale) ou ne correspond pas à un intérêt récurent de l'utilisateur. La considération d'une description collaborative du document selon un utilisateur et son voisinage peut augmenter son score de correspondance et aide à le promouvoir, en particulier lorsque sa requête correspond à un intérêt récurrent chez ses voisins.

En se référant à la figure 5.3, supposons que les utilisateurs 1 et 4 forment le voisinage de l'utilisateur 2. Lorsque cet utilisateur fait une recherche avec la requête q_1, une correspondance « requête-document » qui se base uniquement sur les intérêts individuels de cet utilisateur (la représentation « A »), ne permet pas de localiser le document d_k, contrairement à la représentation hybride « B » qui permet d'étendre la représentation de ce document avec les intérêts de son voisinage (la représentation « C »). Ceci permet de localiser ce document. Cette représentation hybride aide aussi à promouvoir ce document en augmentant son score de correspondance avec la requête lorsque l'appariement qui se base uniquement sur la représentation individuelle de l'utilisateur 2 est faible. C'est le cas avec la requête q_2 qui se compose des jetons : a, b, et h. Cette promotion est rendue possible en augmentant la fréquence d'apparition des jetons a et b de la requête q_2 dans la représentation personnalisée du document lorsqu'elle est étendue avec la représentation « C ». Ceci permet aussi d'étendre la liste des jetons du document qui correspond au contenu de la requête. Cette liste est augmentée avec le jeton « h » qui est absent dans la représentation individuelle « A ». Ceci aide à augmenter son score de correspondance.

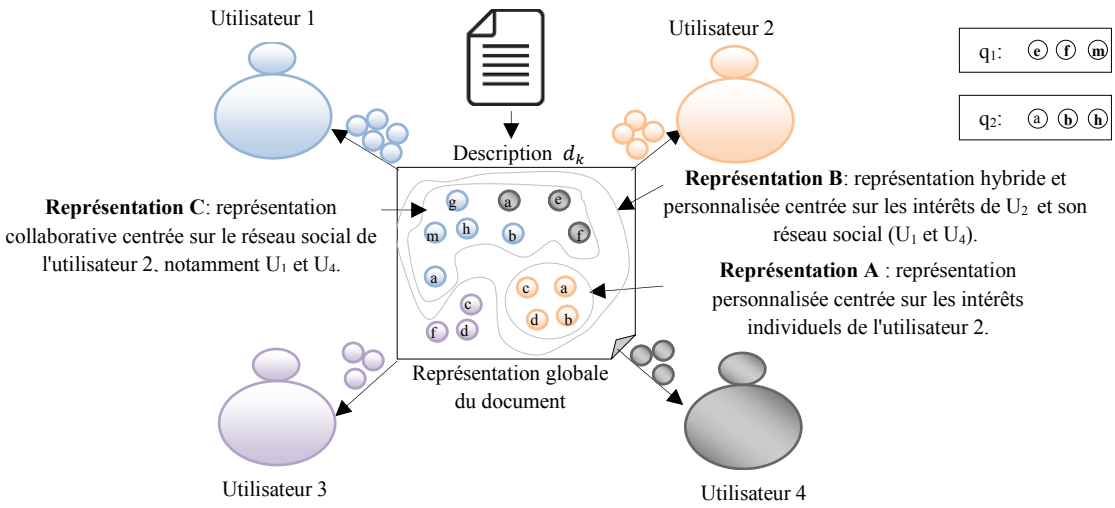


Figure 5. 3. Représentation d'un document centrée sur les intérêts de l'utilisateur et son voisinage

B. Construction du voisinage utilisateur

Ce voisinage est calculé en identifiant pour chaque utilisateur le groupe SFC le plus représentatif de sa recherche actuelle. Contrairement à la prédiction des intérêts, dans une recherche full texte, le besoin de l'utilisateur est explicitement exprimé à travers une requête de recherche. Cette requête n'est pas

toujours claire, mais elle peut aider le système à identifier le besoin en information de l'utilisateur qui aide à son tour à définir son voisinage. Ce voisinage dépend ainsi de la recherche actuelle de l'utilisateur et évolue avec l'évolution de ses taches de recherche. Pour ce faire, le système identifie le sujet de recherche qui couvre le contexte de la requête utilisateur au sein des sujets d'intérêt des utilisateurs stockés dans le système QF. Le système crée alors le profil de la requête (cf. définition 5.2) et le projette sur les sujets du système QF. Le résultat est le sujet qui couvre le contexte de cette requête. Lorsque la requête est ambiguë, le système peut identifier plusieurs sujets qui couvrent son contenu, le système doit alors identifier parmi les sujets résultants celui qui représente le plus les attentes de l'utilisateur en se basant sur son profil. Ce processus sera détaillé dans la section V.2.2 (cf. page 188).

Le sujet sélectionné est utilisé pour identifier le groupe SCF pertinent pour la recherche de l'utilisateur. Il s'agit du groupe SCF qui englobe le sujet identifié. Les utilisateurs qui partagent le même groupe SCF représentent une communauté d'intérêt. Pour chaque utilisateur cible, les utilisateurs appartenant à sa communauté d'intérêt forment son voisinage.

V.2.1.3. Modèle de recherche d'information personnalisée

Ce modèle propose une fonction qui calcule le score de correspondance $RSV(d_k, q, u_j, I)$ d'un document d_k pour une requête q soumise par un utilisateur u_j au sein de l'index multidimensionnel I. Cette fonction compare le contenu de cette requête avec celui du document en fonction de chaque espace de recherche S_i et de l'utilisateur cible u_j. La correspondance se base sur le calcul des occurrences des jetons qui appartiennent à la fois au contenu du document dans l'espace interrogé S_i et à celui de la requête q. On écrit :

$$RSV_s(d_k, q, u_j, Si) = \sum_{tk_i \in (S_{i_{d_k}} \cap q)} PW^{Si}_{d_k, tk_i, u_j} \times W_{q, tk_i} \qquad (5.\ 13)$$

Où $PW^{Si}_{d_k, tk_i, u_j}$ est le score de pondération personnalisé du jeton tk_i dans le document selon l'utilisateur cible u_j. W_{q,tk_i} est le poids de ce jeton dans la requête, souvent égale à 1 ($W_{q,tk_i} = 1$).

Le score de correspondance global d'une requête avec le contenu multidimensionnel d'un document d_k est calculé en fonction des différents scores élémentaires obtenus dans chaque espace Si de recherche (cf. figure 5.4) . On écrit :

$$RSV\ (d_k, q, u_i, I) = \frac{1}{m} \sum_{i=1}^{m} RSV_s\ (d_k, q, u_j, Si) \quad (5.\ 14)$$

Où « m » est le nombre d'espaces de description du document d_k ayant correspondu avec le contenu de la requête q, tel que m ∈ [1,3]

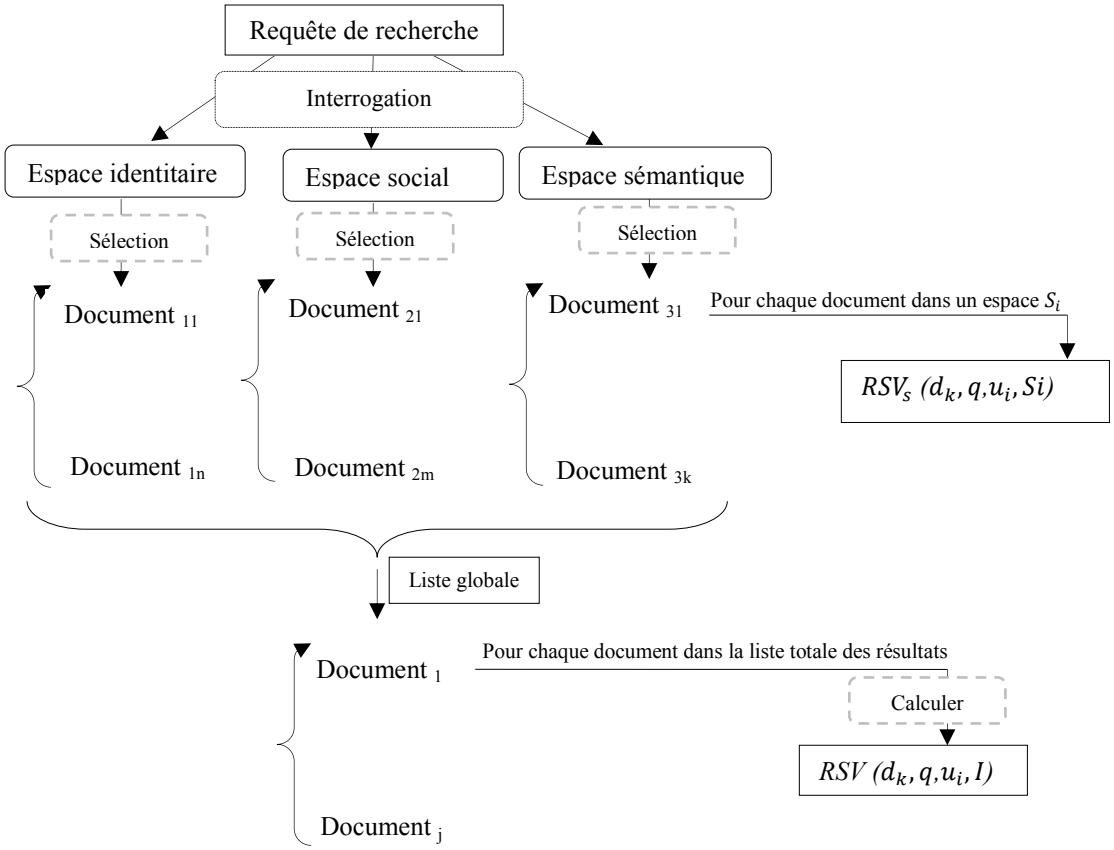


Figure 5. 4. Correspondance de la requête de recherche avec un contenu multidimensionnel

Afin de personnaliser le score de correspondance global d'un document selon les intérêts de l'utilisateur envers les espaces de recherche, un coefficient de pondération noté par $\omega_{S_i}^{u_j}$ est appliquée pour chaque score élémentaire obtenu suite à la correspondance du document avec la requête dans un espace S_i. Ces coefficients permettent de promouvoir les documents provenant d'un ou plusieurs espaces auxquels l'utilisateur s'intéresse le plus. On écrit :

$$RSV\ (d_k, q, u_j, I) = \frac{1}{m} \sum_{i=1}^{m} \omega_{S_i}^{u_j}\ RSV_s(d_k, q, u_j, Si) \quad (5.\ 15)$$

$\omega_{S_i}^{u_j}$ est le degré d'intérêt de l'utilisateur u_j pour l'espace de recherche S_i. Cet espace d'intérêt est nommé dans le modèle du profil utilisateur par la facette d'intérêt et est noté par $fe_i^{u_j}$ (cf. section IV.1.5). Le degré d'intérêt de l'utilisateur pour une facette de données est égal à la proportion de ses objets d'intérêt qui sont liés à cette facette d'intérêt par rapport à tous ses objets d'intérêt. Il est calculé à travers le rapport des deux ensembles de données E_1 et E_2. Où E_1 englobe les d'objets d'intérêt spécifiques de l'utilisateur qui sont liés à cette facette d'intérêt, et E_2 englobe tous les objets d'intérêt spécifiques de cet utilisateur. Ainsi, on écrit :

$$\omega_{S_i}^{u_j} = \frac{card\ (E_1)}{card(E_2)} = \frac{Card\ \left\{obj_k^{fe_i} \in P_{u_j}^{N1}(fe_i)\ pour\ k \in [1, |P_{u_j}^{N1}(fe_i)|]\right\}}{Card\ \left\{\bigcup_{i=1}^{3} obj_k^{fe_i} \in P_{u_j}^{N1}\ pour\ k \in [1, |P_{u_j}^{N1}|]\right\}} \qquad (5.\ 16)$$

Où $P_{u_j}^{N1}$ est le niveau le plus inférieur du profil utilisateur qui englobe ses objets d'intérêt spécifiques, et $P_{u_j}^{N1}(fe_i)$ est le sous-ensemble d'objets qui sont liés à la facette d'intérêt fe_i dans $P_{u_j}^{N1}$.

V.2.1.4. Classement des résultats de recherche à base des facettes d'intérêt de l'utilisateur

Tel que montré dans la figure 5.4, les documents qui correspondent au contenu de la requête de recherche proviennent de trois différents espaces d'indexation. Le score de correspondance de chacun est calculé au sein d'un espace donné. Ainsi, afin de classer tous les documents provenant de l'index I, leurs scores de correspondance sont ajustés selon cet index multidimensionnel et selon l'intérêt de chaque utilisateur pour les facettes de données, et cela en appliquant des coefficients de pondération (cf. équation 5.15). Ce score de correspondance est utilisé pour ordonner la liste des résultats obtenus. Cette liste est renvoyée à l'utilisateur au sein d'une facette de contenu personnalisé.

V.2.1.5. Synthèse

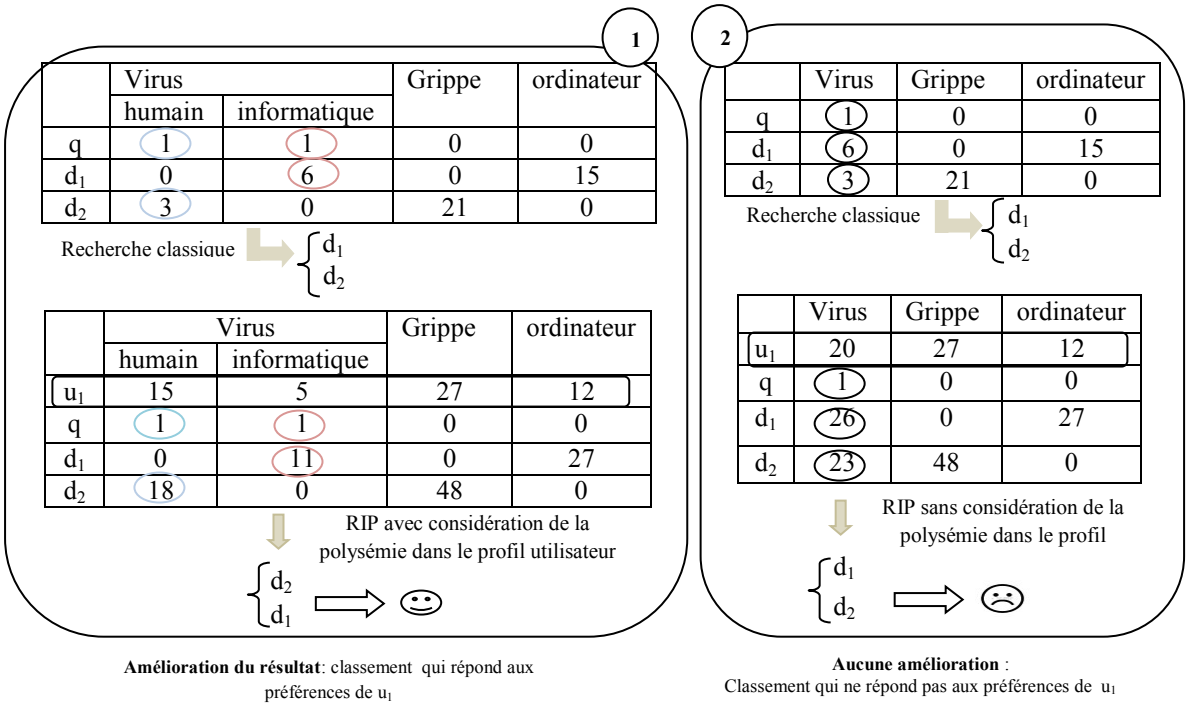
Le nouveau modèle de pondération proposé dans ce chapitre est une amélioration du modèle social BM25FS (Bouhini *et al.* 2013b) (Bouhini 2014). Il tient compte de la polysémie des jetons qui sont exploités dans la description des documents, et de leur représentativité par rapport au contenu de ces documents. Il a été adapté à un système hybride et multidimensionnel. Le modèle social BM25FS exploite les annotations des utilisateurs ainsi que ceux de leurs utilisateurs voisins pour créer un index

personnalisé pour chaque utilisateur. Les majeures différences de notre contribution par rapport à ce modèle se résument par les points suivants :

1. Par rapport à la polysémie des jetons d'enrichissement: dans les travaux de Bouhini et ses collègues, le calcul de la fréquence d'occurrences d'un jeton au sein de la fonction de pondération ne tient pas compte du phénomène de la polysémie. Ainsi, un jeton ayant différentes interprétations et ayant plusieurs occurrences dans le profil de l'utilisateur est considéré comme un même intérêt chez cet utilisateur. De cette façon, lorsque ce jeton est utilisé pour définir une représentation personnalisée des documents sans faire la distinction entre ses différentes interprétations cela peut avoir un impact négatif sur les résultats retournés. Par exemple, lorsqu'un utilisateur utilise le jeton « virus » pour annoter ses documents, l'intérêt de cet utilisateur peut être lié à différentes interprétations (cf. section I.1.4) et deux cas de figure se présentent :

Cas 1 : **l'utilisateur s'intéresse à la fois aux virus humains ainsi qu'aux virus informatiques**. Lorsque le profil de cet utilisateur ne tient pas compte de la polysémie des jetons et est utilisé pour enrichir la représentation des documents dans l'index, cela va augmenter l'occurrence de ce jeton dans le contenu de tous les documents qui sont décrits avec ce jeton puisque le seul critère pris en compte dans ce processus d'enrichissement est une simple correspondance syntaxique entre le contenu du document et celui du profil (Bouhini 2014). De plus, tous ces documents parlant de « virus » seront enrichis avec la même proportion d'occurrence. Ainsi, les résultats de recherche seront similaires à ceux d'un système classique ne tenant pas compte des intérêts de l'utilisateur (cf. figure 5.5 partie 2). Tandis que lorsque la polysémie des jetons est prise en compte dans le profil de l'utilisateur, les documents ne seront pas enrichis de la même manière en particulier lorsque l'utilisateur a une préférence particulière pour l'un des thèmes auxquels est lié le jeton «virus» (cf. figure 5.5 partie 1).

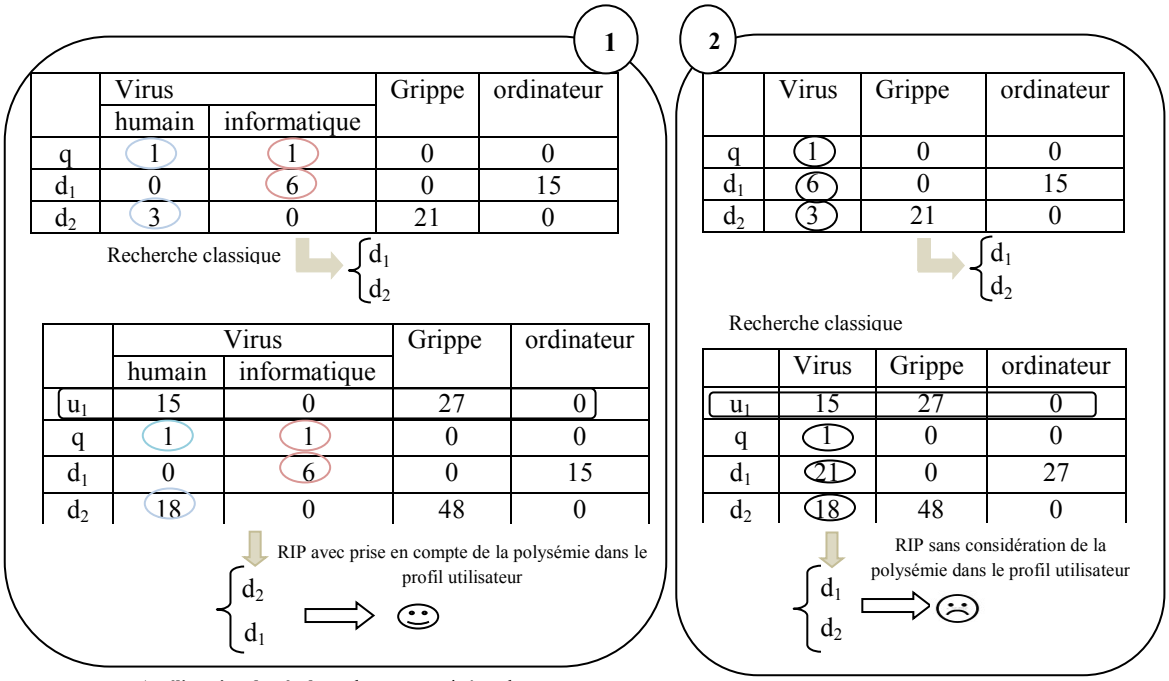
Cas 2 : **l'utilisateur s'intéresse uniquement aux virus humains**. Lorsque le système considère le jeton « virus » comme étant un seul intérêt, celui-ci va être aussi utilisé pour enrichir les documents qui traitent les virus informatiques. Ceci peut avoir un impact sur les résultats qui sont retournés à l'utilisateur. Ces documents peuvent gagner en classement lors d'une recherche avec le jeton « Virus » (cf. figure 5.6 partie 2). Contrairement à cette approche, notre proposition traite cette polysémie et ne permet pas à ces documents d'être influencés lorsqu'ils ne correspondent pas aux intérêts réels de l'utilisateur (cf. figure 5.6 partie 1).



Amélioration du résultat: classement qui répond aux préférences de u_1

Aucune amélioration : Classement qui ne répond pas aux préférences de u_1

Figure 5. 5. Exemple 1 : comparaison entre une RI basée sur un profil contextuel et une RI basée sur un profil non contextuel



Amélioration du résultat: classement qui répond aux préférences de u_1

Aucune amélioration : classement qui ne répond pas aux préférences de u_1

Figure 5. 6. Exemple 2 : comparaison entre une RI basée sur un profil contextuel et une RI basée sur un profil non contextuel

2. Par rapport à l'importance des jetons d'indexation. Lorsque le système exploite les intérêts des utilisateurs pour l'enrichissement du contenu documentaire sans tenir compte de leur représentativité par rapport à ce contenu, cela augmente l'importance des jetons dans les documents même s'ils sont moins représentatifs ou non représentatifs de leur contenu (Bouhini 2014). Par exemple, le cas d'un document qui parle sur le virus de grippe et qui fait référence à d'autres maladies qui sont liées à la grippe, notamment la bronchite ou la toux. Nous considérons ces deux jetons 'bronchite' et 'toux' comme étant moins représentatifs du contenu dudit document. Ainsi, dans notre proposition lorsque le jeton bronchite représente un intérêt pour utilisateur donné, il ne sera pas utilisé pour enrichir le document précité par rapport à cet utilisateur. Il s'agit de sélectionner uniquement les jetons les plus représentatifs d'un document. Cette représentativité est basée sur une pertinence hybride qui combine la pertinence de l'utilisateur et du système (cf. section V.2.1.1).

3. Par rapport à l'évolution des intérêts utilisateur et la flexibilité de son voisinage. Les auteurs dans (Bouhini 2014) exploitent ce qu'ils ont appelé par le profil du voisinage social de l'utilisateur, pour indexer un document. Ce sont les annotations extraites depuis les profils des utilisateurs qui sont similaires à un utilisateur donné. Comme nous l'avons soulevé plus haut, les intérêts des utilisateurs changent au fil du temps et leur voisinage change aussi de son côté. Ainsi, la limitation qui peut être soulevée de la démarche de Bouhini est que le profil social utilisé est considéré comme étant une structure fixe qui ne s'adapte pas à de tels changements. Dans notre cas le voisinage n'est pas prédéfini, il est évalué au fur et à mesure que les intérêts des utilisateurs changent. Ainsi, lors de la soumission d'une requête, le système interroge le contenu des documents en effectuant une recherche personnalisée sur les k-champs qui décrivent ces documents selon l'utilisateur et ses utilisateurs voisins (cf. figure 5.7).

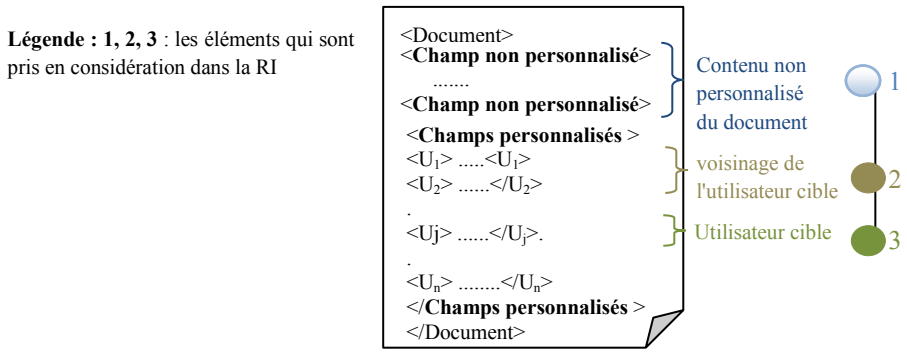


Figure 5. 7. Recherche basée sur une représentation personnalisée du document

4. Par rapport à la structure de l'index. La structure sur laquelle se basent les auteurs pour représenter l'index personnalisé des documents est couteuse en termes d'espace mémoire (Bouhini 2014). Un index personnalisé est construit pour chaque utilisateur. Dans chaque index, les intérêts de l'utilisateur cible notés par Obj_{u_j} ainsi que ceux des autres utilisateurs de son voisinage RS_{u_j} sont également pris en compte en dehors du contenu non personnalisé des documents. Ainsi, si k utilisateurs utilisent le système, et « n » entrées sont nécessaires pour décrire les documents dans un index non personnalisé, alors k*(n + $|Obj_{u_j}|$+| $\cup_{i=1}^{|RS_{u_j}|} Obj_{u_i}$|) entrées sont nécessaires pour décrire ces documents selon les k utilisateurs du système. Contrairement à cette structure, notre proposition utilise pour décrire un document des champs de contenu relatifs aux différents utilisateurs et elle personnalise la recherche sur un sous-ensemble de champs. L'index comprend (| $\cup_{j=1}^{k} Obj_{u_j}$| + n) entrées. Cela réduit considérablement l'espace mémoire par rapport au système de référence (Bouhini 2014). Où |Obj_{u_j} | représente la cardinalité de l'ensemble Obj_{u_j}

5. Par rapport à la diversité d'enrichissement. Contrairement aux travaux de Bouhini (Bouhini 2014) et de Bouadjenek (Bouadjenek *et al.* 2016) qui exploitent soit les annotations des utilisateurs, soit le contenu textuel du document pour décrire un document, notre espace de description est multidimensionnel. Ainsi, la représentation personnalisée des documents que nous proposons dans notre modèle d'enrichissement personnalisé permet aussi d'augmenter l'importance des jetons d'indexation lorsque ces jetons sont liés sémantiquement aux intérêts de l'utilisateur. Prenons l'exemple de la figure 5.8, lorsque les jetons « sofa » et « divan » apparaissent dans l'historique des recherches ou d'annotations de l'utilisateur, les documents qui sont décrits avec « canapé» ou « causeuse» sont aussi pris en considération dans le processus d'enrichissement de l'index orienté utilisateur. Cela est rendu possible grâce à l'espace sémantique qui permet de décrire ces documents avec les jetons qui sont en relations sémantiques avec ceux des autres espaces. Ainsi donc, les documents qui sont décrits avec des jetons qui sont similairement liés aux intérêts de l'utilisateur sont également influencés par ce processus d'enrichissement.

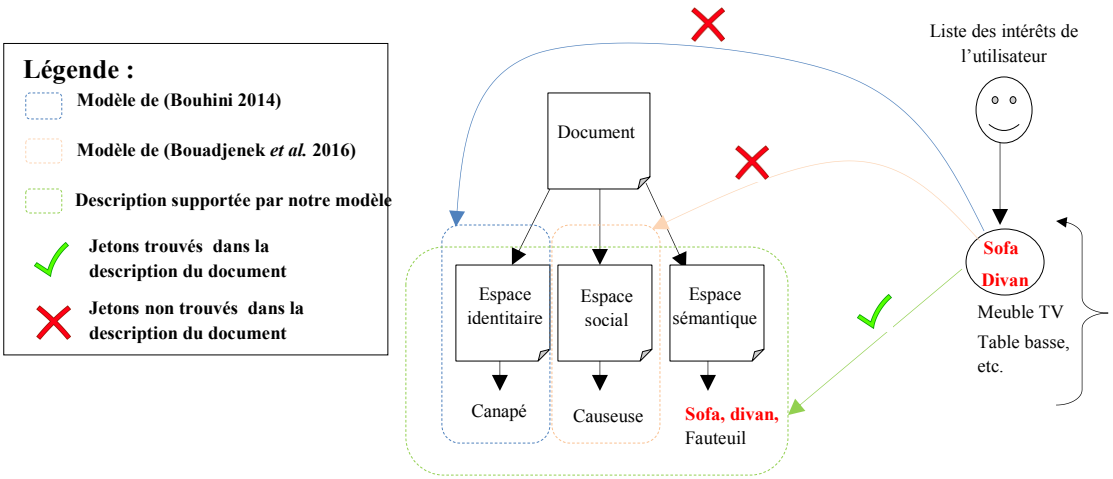


Figure 5. 8. Localisation des intérêts utilisateur dans l'index documentaire

V.2.2. Démarche II: Personnalisation de données basée sur le réordonnancement contextuel des résultats de recherche

Cette démarche consiste à exploiter les données d'intérêt de l'utilisateur uniquement dans le processus de réordonnancement de ses résultats de recherche (Hannech *et al.* 2016a). Il s'agit de retourner les documents qui correspondent au contenu de la requête utilisateur dans l'index documentaire, puis réajuster leurs scores de correspondance selon le contenu du profil utilisateur ainsi que ceux de son voisinage. Cette démarche est basée sur la motivation suivante : étant donné que les utilisateurs ne regardent généralement que les premiers résultats de recherche, le réordonnancement de ces résultats selon leurs intérêts permet de promouvoir les documents qui sont mal classés lorsque la recherche est basée uniquement sur la thématique de la requête. Ce problème se produit fréquemment avec les requêtes ambigües et non précises. L'intégration d'un contexte de recherche aide à augmenter la précision de la recherche. Ce processus est mis en œuvre à travers les étapes suivantes :

Étape 1. Correspondance requête-document : dans cette première étape, le système retourne les documents qui correspondent à la requête de recherche q dans l'index multidimensionnel I. Dans ce cas, le score de correspondance d'un document d_k est évalué selon l'équation 5.17. Ce score est complètement indépendant de l'utilisateur. Le résultat est la liste des documents qui sont sélectionnés vis-à-vis le contenu thématique de la requête, que nous notons par D_q.

$$Score_1\ (d_k, q) = RSV\ (d_k, q, I) = \frac{1}{3}\sum_{i=1}^{3} RSV_s\ (d_k, q, S_i) \quad (5.\ 17)$$

Tel que : $$RSV_s\ (d_k, q, S_i) = \sum_{tk_i \in (S_{i_{d_k}} \cap q)} W_{d_k, tk_i}^{S_i} \times W_{q, tk_i} \quad (5.\ 18)$$

Étape 2. Correspondance document-profil utilisateur : le système évalue ensuite la pertinence de chaque document d_k résultant de l'étape précédente ($d_k \in D_q$) vis-à-vis l'utilisateur cible u. Cela est effectué en déterminant la similarité sémantique du contenu de ce document avec le contenu du profil utilisateur noté par P_u**.** Cette étape est divisée en deux sous-étapes : l'identification du centre d'intérêt utilisateur qui couvre le sujet de sa requête courante dans son profil. Ce centre d'intérêt forme le profil de l'utilisateur à court terme que nous notons par P_u^q. Puis, l'évaluation de la correspondance des documents résultants de l'étape précédente avec le profil temporaire de l'utilisateur P_u^q.

Étape 2.1. Construction du profil à court terme. Pour la construction ce profil, deux cas de figurent se présentent :

- Si la requête courante a été précédemment soumise par l'utilisateur. Le système sélectionne le ou les centres d'intérêt auxquels est liée cette requête dans le profil de cet utilisateur.
- Dans le cas contraire, le système crée le profil de cette requête noté par V_q (cf. définition 5.2), puis le projette sur le profil global de l'utilisateur P_u. Cette projection consiste à calculer la corrélation entre le vecteur conceptuel V_q qui représente le contenu générique de la requête et chaque vecteur V_{IC_i} représentant le contenu conceptuel d'un centre d'intérêt IC_i dans le profil P_u (cf. définition 5.4). En s'appuyant sur le cycle comportemental de l'utilisateur fondé sur les sessions de recherche (cf. section IV.5.4), le système commence d'abord par comparer le profil de la requête avec le centre d'intérêt courant de l'utilisateur auquel est liée sa requête précédente pour savoir si la recherche courante est liée ou non au même besoin précédent. Cette comparaison est effectuée en utilisant la mesure de Kendall qui calcule la corrélation des rangs entre les concepts d'un centre d'intérêt et ceux de la requête dans son profil V_q. Dans ce cas, nous considérons deux vecteurs conceptuels V_1 et V_2 comme étant similaires lorsqu'ils sont liés au même sujet de recherche (cf. définition 4.5). Le résultat final est le centre d'intérêt ayant une corrélation élevée avec le profil de la requête. Cette sélection se base un seuil de corrélation qui sera déterminé expérimentalement.

Lorsque la requête q est ambigüe, elle peut être liée à plusieurs interprétations et son profil est constitué de plusieurs vecteurs conceptuels. Chacun représente un sujet de recherche (cf. Annexe 3). Dans ce cas, le profil temporaire résultant P_u^q peut-être aussi constitué de plusieurs centres d'intérêt, en particulier lorsque cet utilisateur s'est intéressé à plusieurs sujets de recherche auxquels est liée cette requête. Si l'on revient à l'exemple de la section V.2.1.5 d'un utilisateur qui s'est intéressé à la fois aux virus informatiques et humains. Lorsque cet utilisateur soumet à nouveau la requête « q = types of virus » qui présente un contenu ambigu et vaste, deux centres d'intérêt correspondront à cette requête dans le profil de cet utilisateur (cf. figure 5.9). Dans ce cas, afin de personnaliser les résultats de recherche de cet utilisateur, le système sélectionne parmi ces deux centres d'intérêt résultants celui auquel l'utilisateur s'intéresse le plus. Cette sélection se base sur la fraicheur des sujets dans le profil de l'utilisateur et les relations qui les relient les unes aux autres à de hauts niveaux d'abstraction. L'idée est que si le système doit sélectionner le sujet le plus pertinent pour l'utilisateur depuis une liste de sujets notée par SUJ, il se base sur le niveau le plus supérieur qui classifie ces sujets en groupes de SFC pour analyser l'intérêt de cet utilisateur pour les autres sujets qui appartiennent aux mêmes groupes que les sujets dans SUJ.

Pour expliquer mieux cette démarche, prenons l'exemple d'un profil utilisateur illustré par la figure 5.9, et supposons que les interactions de cet utilisateur en termes d'activités de recherche suivent un ordre défini dans le spectre temporel de la figure 5.10. Le résultat de la projection de la requête de recherche q soumise par cet utilisateur est l'ensemble E_1 qui englobe les sujets de recherche sub_2 et sub_4 correspondant au contenu de cette requête q. Pour définir le sujet le plus pertinent parmi ces deux éléments, le système :

- Extrait l'ensemble des sujets qui appartiennent aux mêmes groupes SFC que les sujets dans E_1. Il s'agit de l'ensemble E_2 tel que $E_2 = \{sub_1, sub_3, sub_5\}$ (cf. figure 5.9).
- Sélectionne ensuite le sujet le plus frais dans l'ensemble E. Cet ensemble englobe les sujets de départ avec leurs sujets voisins : $E = (E_1 \cup E_2)$. Pour ce faire, la fraicheur de chaque sujet dans cet ensemble est évaluée. Elle est égale à l'inverse de la période de temps Δt_i écoulée depuis l'instant t' où la dernière activité de recherche A_i liée à ce sujet a été effectuée jusqu'à l'instant t présent. Par exemple, en se référant à la figure 5.10 qui représente la succession chronologique des activités de l'utilisateur et à la figure 5.9 qui représente les relations «$SFC - sujet - activité$», la fraicheur du sujet sub_2 est égale à $\frac{1}{\Delta t_2}$ puisque la dernière activité de recherche effectuée par l'utilisateur u et

appartenant à ce sujet est l'activité A_{14}. Le sujet le plus frais est celui ayant une fraicheur maximale, que nous notons par sub_{Max}. Dans ce cas, le sujet résultant de notre exemple est sub_5.

- En se basant sur le sujet résultant sub_{Max}, le système sélectionne dans l'ensemble de départ E_1 le sujet qui appartient au même groupe SFC_i que sub_{Max}, et le considère comme le sujet pertinent pour l'utilisateur. Comme nous pouvons le voir , malgré que le sujet sub_2 est plus frai que sub_4 puisque $\Delta t_2 < \Delta t_1$, nous considérons que sub_4 est le sujet le plus proche à la requête courante puisqu'il est lié au sujet sub_5 qui est plus frais que les autres sujets dans l'ensemble E. Ce sujet fait référence à un centre d'intérêt dans le profil de l'utilisateur et forme son profil à court terme P_u^q.

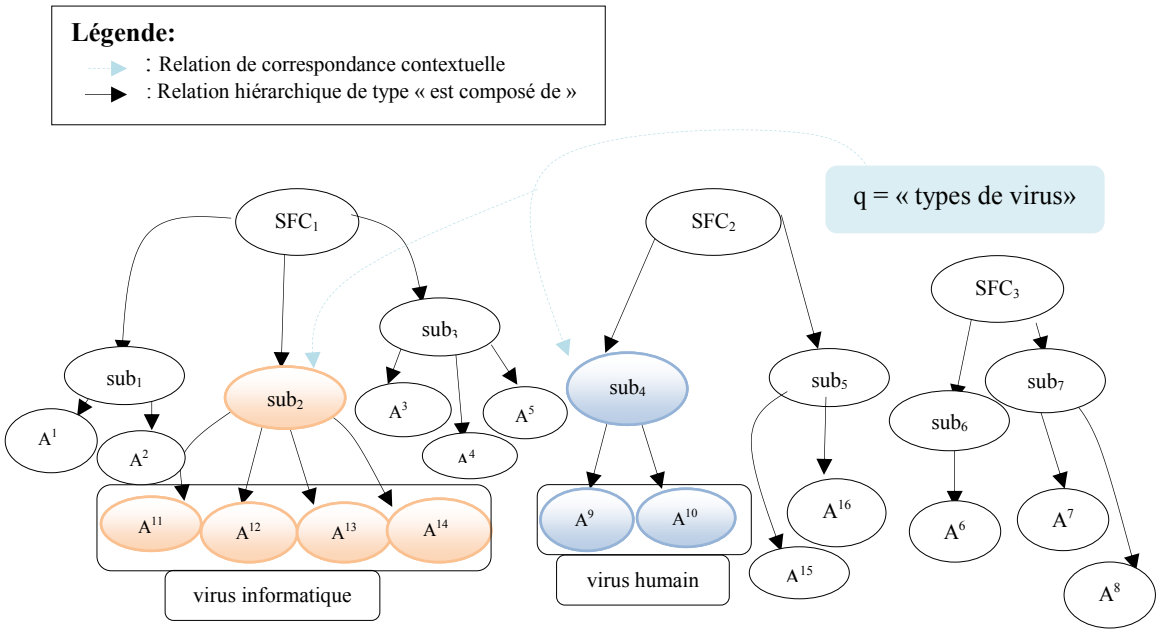


Figure 5. 9. Exemple d'un profil utilisateur

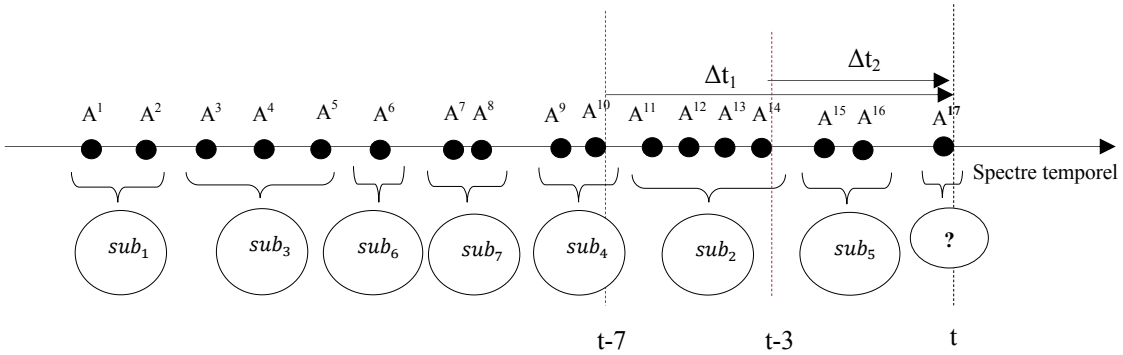


Figure 5. 10. Représentation chronologique des interactions utilisateur

Étape 2.2. Correspondance des documents avec le profil temporaire de l'utilisateur. Le centre d'intérêt IC_i sélectionné est utilisé pour évaluer le score de correspondance de chaque document d_k résultant de l'étape 1($d_k \in D_q$) vis-à-vis l'utilisateur u. Ce score est noté par $Score_2\ (d_k, u)$. Il est égal à la moyenne arithmétique des similarités élémentaires $sim\ (d_k, d_j)$ qui évaluent la similarité du contenu d_k avec le contenu de chaque document d_j dans le centre d'intérêt IC_i. On écrit :

$$Score_2\ (d_k, u) = Sim_{IC}\ (d_k, IC_i\ ) = \frac{1}{\mathrm{N}} \sum_{d_j \in IC_i,\ j=1}^{N} sim_d\ (d_k, d_j) \qquad (5.\ 19)$$

Où N est le nombre de documents dans le centre d'intérêt IC_i. La similarité élémentaire $sim_{\ e}(d_k, d_j)$ est la similarité entre le document cible d_k et le document d_j appartenant au centre d'intérêt IC_i. Elle est évaluée par le calcul du cosinus de l'angle entre leurs vecteurs respectifs $V_{d_k}\ et\ V_{d_j}$ (cf. définition 5.1).

Définition 5.1. Représentation vectorielle du contenu documentaire. Il s'agit du vecteur des termes pondérés constituant le contenu identitaire d'un document. Ces termes sont pondérés par le modèle BM25F. Cette pondération est définie dans l'équation 5.1.

Définition 5.2. **Profil de la requête de recherche**. Il s'agit de la liste des concepts qui représentent les domaines d'intérêt correspondant au contenu de la requête q. Ces concepts sont obtenus en projetant le contenu de cette requête sur une ontologie topique (cf. définition 4.1) que nous notons par O_D. Ce profil est modélisé par un vecteur de concepts pondérés. Chaque pondération représente le résultat de la projection de q sur O_D en termes de similarité sémantique. On écrit :

$$V_q = \{(C_1, score_1), \ldots, (C_j, score_j), .., (C_k, score_k)\ \} \qquad (5.\ 20)$$

Tel que :

$$score_j = \text{cosinus}\ (\vec{q}, \overrightarrow{C_O}\ ) \qquad (5.\ 21)$$

Où C_o est un concept correspondant à un domaine d'intérêt dans l'ontologie O_D et $\overrightarrow{C_O}$ est la représentation vectorielle de son contenu informationnel (cf. définition 5.1). Afin de sélectionner les concepts les plus pertinents au contenu de la requête, un graphe de requête est construit permettant de garder que les concepts qui sont liés sémantiquement les uns aux autres (cf. définition 5.3). Les concepts qui restent déconnectés du graphe sont éliminés. Cette démarche est aussi utile lorsque la requête est liée à plusieurs

sujets de recherche. Le résultat est un ensemble de graphes chacun représente une composante de concepts connexes et forme un profil d'un sujet de recherche.

Définition 5.3. Un graphe de requête G_q^t consiste à relier sémantiquement l'ensemble des k-concepts les plus représentatifs de la requête q au sein d'un graphe conceptuel en exploitant l'ontologie de référence O_D. Le processus de construction de ce graphe est détaillé dans l'annexe 4.

Définition 5.4. Représentation vectorielle d'un centre d'intérêt. Il s'agit du vecteur conceptuel pondéré qui représente l'ensemble des domaines d'intérêt qui forment un centre d'intérêt de l'utilisateur (cf. définition 4.7). Les pondérations représentent les degrés d'intérêt de l'utilisateur pour ces domaines d'intérêt. Ils sont accumulés durant ses activités de recherche (cf. section IV.5.4.2).

Étape 3 : Le troisième score qui évalue la pertinence d'un document d_k ($d_k \in D_q$) représente sa correspondance vis-à-vis les profils des autres utilisateurs qui forment le voisinage de l'utilisateur cible que nous notons par RS_u. De la même manière présentée dans la section V.2.1.2.4, ce voisinage est calculé. Le système l'exploite pour extraire les données d'intérêt de ces utilisateurs voisins qui correspondent au besoin informationnel courant de l'utilisateur cible. Il consiste à extraire depuis leur profils, l'ensemble des centres d'intérêt qui correspondent au sujet d'intérêt de l'utilisateur cible identifié dans l'étape 2 et stocké dans son profil temporaire, puis l'exploiter dans l'évaluation de la pertinence des documents D_q. Cet ensemble est noté par IC_{RS_u}. On écrit :

$$Score_3\ (d_k, RS_u) = sim_{RS}\ (d_k, IC_{RS_u}) = \frac{1}{\mathrm{k}} \sum_{\mathrm{IC_i} \in \mathrm{IC_{RS_u}}, \mathrm{i}=1}^{k} sim_{IC}\ (d_k, IC_i) \qquad (5.\ 22)$$

Où k est le nombre de centres d'intérêt qui existent dans l'ensemble IC_{RS_u}. L'évaluation de la similarité élémentaire $sim\ (d_k, IC_i)$ suit la même formule qui a été définie dans l'équation 5.19.

La valeur de similarité obtenue avec le 3eme score $Score_3\ (d_k, RS_u)$ est particulièrement élevée lorsque le nombre des centres d'intérêt IC_i dans l'ensemble $\mathrm{IC_{RS_u}}$, ayant une grande similitude avec d_k ,est grand. Dans ce cas, d_k est considéré comme pertinent par plusieurs utilisateurs voisins RS_u. Cela augmente son importance pour l'utilisateur cible u.

Étape 4. Une fois les trois scores de correspondance du document d_k sont évalués, le score global évaluant sa correspondance personnalisée avec la recherche cible est évalué à travers la combinaison linéaire de ces scores élémentaires. Ce score global est utilisé pour ordonner les documents. On écrit:

$$Score_G\ (d_k, q, u) = \alpha\ (score_1\ (d_k, q))\ + \beta\ ((score_2\ (d_k, \mathrm{u})) + \omega(score\ _3(d_k, RS_u)) \qquad (5.\ 23)$$

Où α, β, ω sont des coefficients de pondération qui représentent les degrés d'importance de chaque score élémentaire dans l'évaluation de la pertinence globale du document.

V.2.2.1. Proposition de nouveaux documents pour l'utilisateur

Nous avons proposé dans la section précédente une fonction linéaire qui permet de fournir à un utilisateur des résultats personnalisés en réponse à sa requête de recherche. Cela est effectué en réordonnant les documents qui résultent de cette recherche selon les intérêts de cet utilisateur qui couvrent le contexte de cette requête et ceux des autres utilisateurs voisins. Dans cette section, nous proposons une méthode qui permet d'offrir une liste de documents qui n'ont pas encore été considérés par cet utilisateur, notamment lorsque sa requête de recherche courante est liée à un centre d'intérêt dans son profil. Cela permet de proposer de nouveaux résultats qui contribuent à l'enrichissement de son profil. Ce processus se base sur les expériences des autres utilisateurs en vue de prédire l'intérêt de cet utilisateur pour les nouveaux documents proposés. Cette prédiction est illustrée à travers la figure 5.12 et se résume par les étapes suivantes :

Étape 1 : Le système construit le profil à court terme de l'utilisateur cible u (profil temporaire).

Étape 2 : Il calcule ensuite le voisinage de l'utilisateur cible noté par RS_u. Cette étape se base sur deux sous taches : la première consiste à calculer le groupe d'intérêt de l'utilisateur qui permet d'obtenir un cluster d'utilisateurs qui portent leurs intérêts pour le même groupe SFC (cf. section V.2.1.2.4), puis extraire depuis le cluster sélectionné les utilisateurs les plus similaires. Cela aide à réduire l'espace des calculs en identifiant d'abord le cluster d'utilisateurs les plus similaires selon un niveau plus générique (les groupes de SFC) puis identifier les utilisateurs voisins selon une similarité plus spécifique (cf. figure 5.11). Cette similarité spécifique se base sur la comparaison de leurs documents d'intérêts. Cette similarité spécifique se base sur la motivation suivante : les utilisateurs qui ont consulté plus de

documents en commun avec l'utilisateur cible sont plus pertinents pour prédire l'intérêt de cet utilisateur pour un nouveau document.

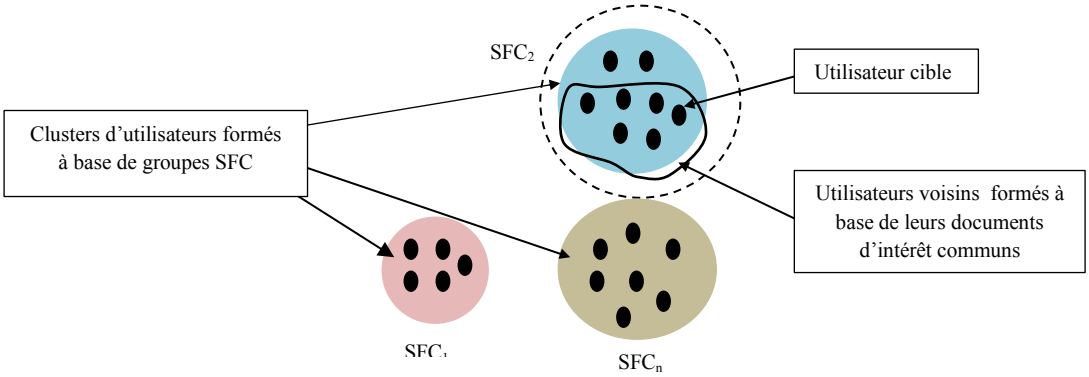


Figure 5. 11. Sélection des utilisateurs voisins depuis un cluster d'utilisateurs

Étape 3. Les sujets d'intérêt des utilisateurs voisins RS_u sont préparés. Puis, le système sélectionne depuis la liste des sujets préparée, le sujet de recherche qui correspond à la requête de l'utilisateur. Cela est effectué en projetant le profil temporaire de cet utilisateur P_u^q sur cette liste de sujets. Cette projection est identique à la projection expliquée dans la section suivante. Le résultat est le sujet sub_q qui couvre le contexte de la recherche utilisateur.

Étape 4: À partir du sujet résultant, une matrice «utilisateur-document» est élaborée. Elle représente l'utilisateur cible u et ses utilisateurs voisins $RS_u^{sub_q}$ avec leurs documents d'intérêt qui couvrent le sujet de recherche sub_q auquel est liée la requête q. Chaque cellule dans la matrice représente la popularité d'un document d_k par rapport à un utilisateur u_j que nous notons par $Pop\ (d_k, u_j, sub_q)$. Cette popularité est estimée en termes de fréquence d'apparition de ce document dans le profil de l'utilisateur u_j par rapport aux fréquences des autres documents. Elle est formulée comme suit**:**

$$Pop\ (d_k, u_j, sub_q) = \begin{cases} \frac{fr(d_k, u_j, sub_q)}{fr(ArgMaxfr(d_i, u_j))}, \text{ si } ArgMaxfr(d_i u_j) \neq \emptyset \\ 1, \text{ sinon} \end{cases} \quad 5.\ 24$$

Où $fr(d_k, u_j, sub_q)$ est la fréquence d'apparition d'un document d_k lié au sujet de recherche sub_q dans le profil de l'utilisateur u_j. $ArgMaxfr(d_i, u_j)$ est la fonction qui permet de retourner un document d_i appartenant au profil de l'utilisateur u_j ayant une fréquence d'apparition maximale.

Étape 5. La matrice construite à l'étape précédente est utilisée pour prédire l'intérêt que l'utilisateur cible u peut porter aux documents de ses voisins qu'il n'a pas encore considérés. Ainsi, le degré d'intérêt $score(u, d_k)$ de cet utilisateur pour un document d_k est égal à la moyenne pondérée des popularités de ce document par rapport aux autres utilisateurs voisins $RS_{u_i}^{sub_q}$ dans la matrice.

$$score(u, d_k) = \frac{\Sigma_{u_j \in RS_{u_i}^{Sub}, j=1}^{n} w_j * Pop\,(d_k, u_j)}{\Sigma_{j=1}^{n} Sim\,(u, u_j)} \qquad 5.\,25$$

La pondération w_j attribuée à chaque valeur de popularité $Pop\,(d_k, u_j)$ d'un document d_k à prédire par rapport à l'utilisateur voisin u_j, est déterminée par le degré de similarité entre cet utilisateur et l'utilisateur cible u. Cette similarité est évaluée à travers le calcul de cosinus entre les vecteurs pondérés qui sont associés aux utilisateurs depuis la matrice de popularités (cf. figure 5.12). Ces vecteurs représentant les intérêts des utilisateurs pour les différents documents. On écrit:

$$w_j = sim\,(u, u_j) = \text{cosinus}(\vec{u}, \vec{u_j}) \qquad 5.\,26$$

Tel que :

$$\vec{u} = (Pop_1(u, d_1), .., Pop_k(u, d_k)) \qquad 5.\,27$$

Cette mesure de cosinus est choisie pour les mêmes motivations présentées dans la section IV.5.4.4.1 (cf. page 162).

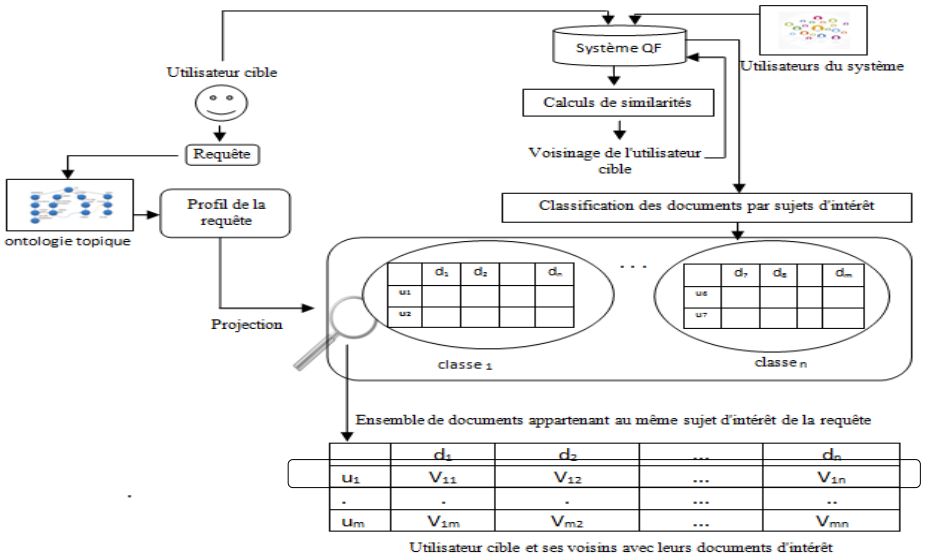


Figure 5. 12. Processus de prédiction de nouveaux documents pour un utilisateur

V. 3. Analyse et conclusion

Ce chapitre a présenté deux modèles qui permettent de personnaliser les résultats de recherche des utilisateurs ayant divers et différents besoins. Cette personnalisation est effectuée pour chaque utilisateur cible en fonction de ses intérêts ainsi que ceux des autres utilisateurs similaires. Pour cela, une technique automatique de création de communautés d'intérêts a été proposée. Elle permet d'affecter l'utilisateur à un groupe d'utilisateurs les plus similaires à ses intérêts actuels, appelé la communauté d'intérêt. Cette communauté est dynamique, elle change au fur et à mesure que les intérêts de l'utilisateur changent. Pour cela, des critères d'importance sont intégrés, à savoir la fréquence et la fraicheur des données. Cela favorise la sélection de la communauté d'intérêt la plus représentative des intérêts de chaque utilisateur. De cette façon, la technique proposée s'adapte à l'évolution des intérêts des utilisateurs au fil du temps et à leur récurrence.

Les deux propositions se distinguent l'une de l'autre principalement par i) le niveau d'intégration des données utilisateurs dans le SRI, par ii) le type de données qui sont exploitées dans le processus de personnalisation, ainsi que par iii) la technique adoptée pour cette intégration. Dans la première proposition, les données d'intérêt sont intégrées dans le processus d'indexation des documents. Cette

démarche ne nécessite pas une couche de traitement supplémentaire lors d'une recherche effectuée par l'utilisateur, car l'intégration est effectuée au préalable dans l'univers d'indexation et le processus de recherche interroge directement l'index étendu. Elle est par contre coûteuse en espace mémoire et délicate à mettre en œuvre dans le cas des corpus documentaires volumineux avec un grand nombre d'utilisateurs. C'est la raison pour laquelle nous avons proposé une autre technique où la personnalisation intervient uniquement au niveau du réordonnancement des résultats. Celle-ci exploite une mesure de corrélation contextuelle entre la requête utilisateur et un centre d'intérêt cible. Ceci permet de cibler directement les données dans le profil couvrant le sujet de la recherche cible. Contrairement aux approches qui se basent sur une similarité sémantique entre la requête de l'utilisateur cible et ses objets d'intérêt stockés dans son profil (Anuradha R. Kale June-2013) (Aicha Aggoune 2016), notre approche se base sur un aspect plus générique, notamment les sujets d'intérêt pour cibler l'ensemble de données répondant à la recherche utilisateur, et les sujets fortement connexes pour définir les utilisateurs voisins.

Nous avons proposé une technique qui fait face à l'ambigüité des requêtes de recherche, en particulier lorsque le profil de l'utilisateur cible ne permet pas seul de résoudre ce problème d'ambigüité. Cela se produit lorsque plusieurs centres d'intérêt qui couvrent la requête de l'utilisateur sont identifiés dans son profil. Cette technique se base à la fois sur la fraicheur des données dans le profil de l'utilisateur et sur les relations qui relient les sujets les unes aux autres au niveau supérieur du profil. Le tableau 5.2 illustre une comparaison entre quelques principales méthodes de la recherche personnalisée proposées dans la littérature avec nos deux modèles proposés dans ce chapitre.

.

Tableau 5. 2. Analyse comparative entre quelques principaux modèles de personnalisation de la littérature et nos modèles

Type d'intégration → / Niveau d'intégration / Critères de comparaison / Critère de comparaison	Au niveau de la requête [1] (Bouhini, Géry et al. 2016) [2] (Zhou, Wu et al. 2017)	Au niveau de la représentation des documents Dans l'index documentaire [1] (Bouhini, Géry et al. 2013) [2] (Aicha Aggoune 2016) [3] Notre modèle	Au niveau de la représentation des documents Au niveau du traitement de la requête (Bouadjenek, Hacid et al. 2016)	Dans le réordonnancement des résultats [1](Anuradha R. Kale June-2013)[2] [2](Venkataraman and Ravichandran 2014)	Dans le réordonnancement des résultats Notre approche (Hannech, Adda et al. 2016)
Couche de traitement requise lors de la recherche	**Oui** [1]Détection des termes pertinents pour l'enrichissement de la requête de recherche. [2]Récupération du contenu documentaire nécessaire pour l'enrichissement du profil utilisateur. [2]Calcul des poids des termes à partir d'un profil d'utilisateur correspondant à la recherche courante.	**Non** [1,2,3] Indexation faite au préalable.	**Oui** Repose sur la factorisation matricielle pour calculer durant le traitement d'une requête de recherche, la représentation des documents correspondant à cette requête.	**Oui** Identification du contenu pertinent qui sera exploité pour le réordonnancement des résultats : [1]Calcul de similarité sémantique entre le profil utilisateur et la requête courante. [2]Correspondance entre les caractéristiques générales des documents web et des documents d'intérêt de l'utilisateur.	**Oui** Identification du contenu pertinent pour la personnalisation de données : i. Identification du contexte de la tache de recherche courante. ii. Projection du contexte de la tache courante sur le profil de l'utilisateur.
Techniques d'enrichissement et stratégies de recherche	[1] Les aspects contextuel et temporel ne sont pas considérés dans la représentation des données d'intérêt de l'utilisateur. [2] Aspect contextuel supporté par une technique de plongement lexical (word embedding) [1,2] Aspect non évolutif des besoins utilisateur et aspect collaboratif non évolutif.	[1,2] Pas d'aspect temporel dans la représentation des données d'intérêt exploités ni dans la recherche [1,2] Pas d'aspect contextuel dans la représentation des données d'intérêt ni dans la recherche. [2] Aspect sémantique dans la représentation des documents. [3] Intégration de l'aspect contextuel et temporel dans la représentation des données.	Les aspects temporel et contextuel des données d'intérêt ne sont pas supportés.		i. Aspect contextuel pris en compte dans la représentation des données d'intérêt de l'utilisateur et dans le processus de RI. ii. Aspect temporel considéré dans la représentation de données d'intérêts des utilisateurs et dans la création des communautés d'utilisateurs. iii. Aspect évolutif de données.
Espace mémoire supplémentaire requis pour la personnalisation en dehors de la base des profils utilisateurs	[1]**Non** Les données sont utilisées directement dans l'enrichissement de la requête. [2]**Oui** Représentation de plongement lexical et de domaines pondérés pour l'enrichissement du profil utilisateur.	**Oui** [1,2,3] Extension de l'index avec les données d'intérêt des utilisateurs.	**Non** La représentation personnalisée des documents est effectuée à la volée lors de la recherche.	**Non** [1,2] Les données d'intérêt des utilisateurs sont exploitées à la volée dans le réordonnancement des résultats de l'utilisateur.	Les données d'intérêt des utilisateurs sont exploitées à la volée dans le réordonnancement des résultats de l'utilisateur.
Critères sur laquelle se base la pertinence des données d'intérêt exploitées	[1,2] La pertinence des données est considérée uniquement par rapport à l'utilisateur cible. [3] Pertinence hybride : pertinence système et utilisateur				
Type de données d'intérêt utilisateur exploitées dans le processus de personnalisation	[1]Un seul niveau de représentation : les étiquettes d'annotation de l'utilisateur. [2]Plusieurs niveaux de représentation : les documents, les annotations et les domaines d'intérêt.	[1]Un seul niveau de représentation : les étiquettes d'annotation de l'utilisateur. [2]Un seul niveau de représentation: les concepts relatifs aux documents d'intérêt de l'utilisateur. [3]Plusieurs niveaux de représentation: **Données concrètes** : requêtes de recherche, étiquettes d'annotation. **Données abstraites** : sujets d'intérêt et groupes de sujets connexes.	Un seul niveau de représentation: les étiquettes d'annotation de l'utilisateur	[1,2] Un seul niveau de représentation : les documents d'intérêt de l'utilisateur.	Plusieurs niveaux de représentation: **Données concrètes** : les documents, les facettes d'intérêt. **Données abstraites** : sujets d'intérêt et groupes de sujets connexes.

Chapitre 6 : Stratégie de recommandation à démarrage à froid basée sur une carte de communautés et l'identification d'utilisateurs centraux

VI.1. Introduction

Le domaine de la personnalisation de données est venu proposer différentes techniques de filtrage d'information en vue d'améliorer la recherche de l'utilisateur et/ou de lui recommander de nouvelles données d'intérêt qui peuvent l'intéresser et enrichir davantage son profil. Ces techniques se basent principalement sur l'exploitation des informations relatives à cet utilisateur et/ou aux autres utilisateurs de sa communauté d'intérêt, appelée aussi son voisinage. Cependant, ces informations sont indisponibles dans certaines situations, en particulier, lorsque l'utilisateur cible utilise le système pour la première fois. Le système se retrouve incapable de lui suggérer du contenu personnalisé ou de l'assigner à une communauté d'utilisateurs sur laquelle une technique de filtrage collaboratif s'appuie généralement pour lui fournir des données personnalisées. Ainsi, les données proposées à cet utilisateur peuvent être loin de ses attentes et ses intérêts réels. Ce problème est connu sous le nom de démarrage à froid d'un nouvel utilisateur. Dans ce chapitre, nous proposons une approche visant à pallier ce problème par la proposition d'une carte de communautés d'utilisateurs, puis l'identification des utilisateurs importants au sein de chaque communauté comme issue de recommandation (Hannech *et al.* 2016b). Cette importance est déterminée en termes de plusieurs critères qui visent à offrir une meilleure exploration du contenu du système.

VI.2. Synthèse

Pour remédier au problème soulevé dans ce chapitre, plusieurs approches ont été suggérées dans la littérature (cf. section II.2.3.1.6). Ces propositions peuvent être résumées en trois catégories:

1. Les approches qui exploitent des sources d'information externes pour compenser le manque d'information sur les intérêts de l'utilisateur, telles que les données personnelles et démographiques de

l'utilisateur. Le recours à l'utilisation des ontologies, des modèles décisionnels et des règles d'association qui exploitent de leur côté comme données d'entrée les informations statiques relatives à l'utilisateur, etc. La majeure limitation avec de telles propositions est que telles informations sur l'utilisateur ne sont pas toujours disponibles.

2. Les approches qui proposent de rapprocher les utilisateurs les uns aux autres à travers différents critères de similarité en dehors de leurs données d'activités, telles que les relations de confiance (ex. relation d'amitié et de suivi qui relient les utilisateurs dans les réseaux de connaissances ou scientifiques), le regroupement géographique qui permet de regrouper les utilisateurs à base de leurs données géographiques, etc. La majeure limitation avec de telles approches réside dans le fait qu'elles se basent sur les relations qui existent déjà entre les utilisateurs dans le réseau social. Celles-ci ne sont pas toujours présentes en particulier lorsque l'utilisateur cible est nouveau et n'est lié à aucun autre utilisateur ou communauté du système.

3. La proposition d'un processus d'entrevue initiale qui interroge progressivement les nouveaux utilisateurs sur leurs intérêts et préférences (Benhamdi *et al.* 2017). Cette phase interrogative peut être intéressante pour certains utilisateurs comme elle peut être ennuyeuse pour d'autres. Cela peut présenter une lourde tâche pour l'utilisateur.

Contrairement à ces propositions, l'approche proposée dans ce chapitre n'a recours à aucune source d'information relative à l'utilisateur cible, telles que ses données personnelles ou ses utilisateurs voisins. Elle suppose que l'utilisateur n'est lié à aucune communauté d'intérêt et ne lui demande de fournir aucune information statique.

VI.3. Idée générale

La contribution consiste en une nouvelle approche de recommandation qui propose une carte de communautés d'utilisateurs, appelées aussi les groupes d'intérêts. Elle organise les réseaux sociaux en plusieurs niveaux d'intérêt où chaque niveau permet de regrouper les utilisateurs en communautés d'intérêts formées sur la base d'un critère donné. Puis, elle identifie les meilleurs points d'entrée de chaque communauté qui aident les nouveaux utilisateurs à explorer facilement les données d'intérêt des autres utilisateurs réguliers du système. Chaque communauté peut être liée à un domaine d'intérêt ou à un

sujet de recherche ou à un groupe de SFC selon le niveau d'exploration dans la carte. Cette carte permet au système d'apprendre les intérêts d'un nouvel utilisateur pour construire rapidement son profil en affinant ses groupes d'intérêt qui peuvent être utiles pour alimenter un système d'inférence et bénéficier des données de recommandation. Pour ce faire, différents critères de regroupement sont d'abord définis pour la formation de communautés. Par la suite, différentes mesures d'analyse des réseaux sociaux sont introduites afin de caractériser le rôle et l'importance de chaque individu dans ce réseau en vue d'identifier un individu « responsable » de chaque communauté. Celui-ci représente le point d'entrée au contenu de sa communauté (cf. figure 6.1). L'importance et les rôles des individus dans les réseaux sociaux évoluent au fil du temps, ainsi l'individu responsable est considéré aussi comme étant dynamique.

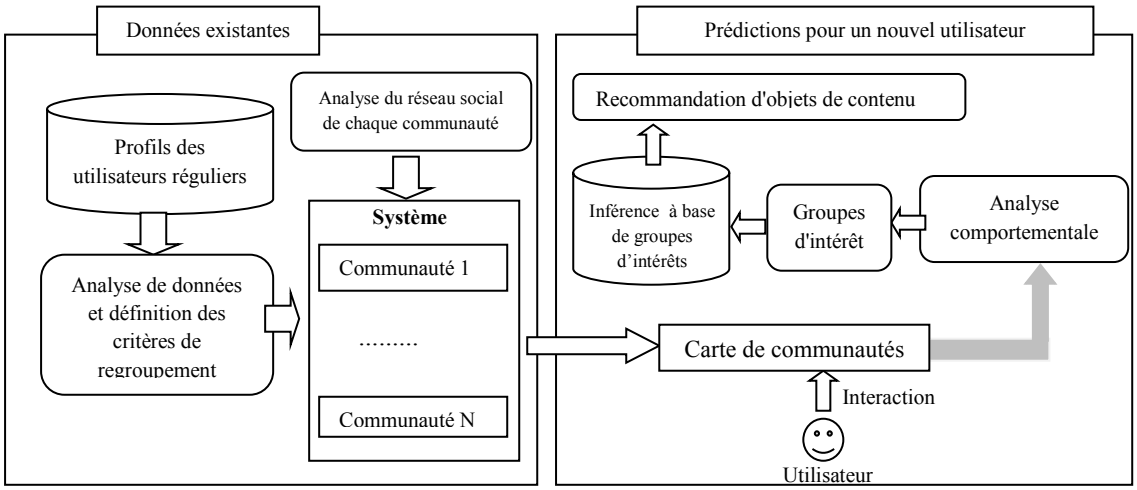


Figure 6. 1. Architecture générale du processus de recommandation pour un nouvel utilisateur

La diffusion de l'information dans un réseau est loin d'être un de nos objectifs, mais peut être considérée comme une solution pour notre problème. Il consiste à choisir un individu avec lequel on peut atteindre rapidement les autres individus qui possèdent dans leurs profils des informations pertinentes pour un nouvel utilisateur. La position de cet individu dans le réseau social doit garantir une exploration rapide et maximale du contenu dans la communauté ou du système. La figure 6.2 montre un exemple de flux d'information. Si la distribution d'information commence par le nœud «A», la diffusion à travers le réseau social s'effectue en 2 sauts. En revanche, une distribution qui commence par le nœud « B » fait l'objet de 4 sauts. Le point «A» représente donc l'individu le plus proche des autres membres du réseau. Ainsi, certains nœuds jouent un rôle plus important que d'autres, ils ont une meilleure capacité à diffuser

l'information via le réseau social. Ces individus peuvent être utiles pour aider les nouveaux utilisateurs dans leurs tâches d'explorations du contenu dans les SRIs en général et dans les réseaux sociaux en particulier. Ce critère d'importance est donc un des facteurs qui sont considérés dans la définition de notre mesure d'importance globale. Nous pensons qu'un individu est capable de représenter une communauté lorsque sa position dans la communauté peut atteindre un maximum d'individus en moins d'étapes et qui possède de solides (fructueuses) relations avec ses contacts voisins (directs et indirects). De là, le score d'importance global d'un membre dans le réseau est défini par un vecteur de mesures au lieu d'une seule mesure. Dans la section suivante, nous présentons les principaux concepts qui sont liés à l'approche proposée.

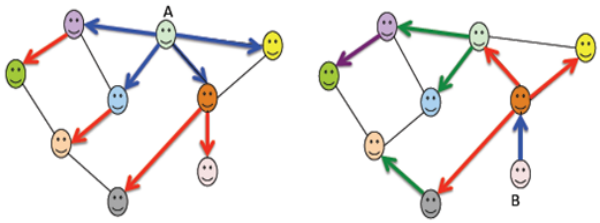


Figure 6. 2. Exemple de flux d'information dans un réseau social (Sarr *et al.* 2012)

VI.4. Concepts de base

Un réseau social est un outil que les utilisateurs utilisent de plus en plus pour communiquer et partager de l'information. Il décrit une structure dynamique généralement représentée sous un graphe G = (V, E), d'un ensemble d'individus représenté par les nœuds V, ils sont liés les uns aux autres par des relations représentées par les arêtes E. Ces relations dépendent du contexte de l'application. Elles peuvent être des relations d'amitié dans le cas d'un réseau de connaissances tel que Facebook et Twitter, etc., des relations de citations dans un réseau de publication scientifique tel que Researchgate et Academia, des liens de connexions physiques ou logiques dans un réseau informatique, etc. Ces réseaux sociaux possèdent plusieurs caractéristiques très intéressantes qui méritent d'être étudiées et analysées.

VI.4.1. Analyse des réseaux sociaux

L'analyse des réseaux sociaux peut être effectuée pour divers objectifs. On retrouve i) les tâches axées sur la communauté pour l'identification des communautés, ii) les tâches orientées vers la structure

du réseau pour la prédiction et la détection des liens cachés ou alarmants, et l'étude de l'évolution du réseau. Et enfin, ii) les taches orientées nœuds pour caractériser le rôle et/ou la position d'une entité dans le réseau. En effet, chaque nœud joue un rôle plus important dans le réseau. Si on considère à titre d'exemple le graphe de la figue 6.3, on remarque que la disparition du nœud « B » n'affecte pas l'exploration d'information dans le graphe, tandis que la disparition du nœud «A» engendre un grand problème d'exploration puisque plusieurs nœuds se retrouvent isolés après sa disparition.

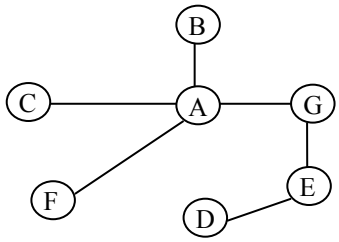


Figure 6. 3. Graphe d'un réseau social

VI.4.2. Mesures d'importance

Dans le but de quantifier la notion d'importance d'un nœud, des chercheurs ont proposé plusieurs définitions connues sous le nom des mesures de centralité (Sarr *et al.* 2012). Celles qui sont fréquemment utilisées sont: la centralité de degré, la centralité d'intermédiarité, et la centralité de proximité.

La centralité de degré d'un nœud individuel fournit le nombre de ses liens directs. La centralité de degré normalisée d'un nœud dans un réseau social G de taille n, est calculée comme suit :

$$C_D^G(i) = \frac{d(i)}{n-1} \qquad (6.\ 1)$$

Pour un nœud i où d(i) est le degré qui représente le nombre de liens directs.

La centralité d'intermédiarité d'un nœud exprime le degré de contrôle qu'il a sur les interactions des autres nœuds du réseau. En d'autres termes, c'est le nombre de nœuds auxquels un nœud i est connecté de façon indirecte, via ses liens directs. La centralité d'intermédiarité normalisée est calculée comme suit :

$$\mathrm{C_I^G(i)} = \frac{2 * \sum_{i \neq j \neq k} \frac{P_{jk}\ (i)}{P_{jk}}}{(n-1)(n-2)} \qquad (6.\ 2)$$

pour un nœud i où P_{jk} est la longueur du chemin le plus court entre les nœuds j et k et P_{jk} (i) est la longueur du chemin entre les nœuds j et k en passant par le nœud i.

La centralité de proximité d'un nœud individuel indique comment un nœud est proche des autres nœuds dans le réseau et donc la façon dont l'information circule rapidement à partir d'un nœud jusqu'aux autres nœuds qui sont accessibles dans le réseau. La centralité de proximité normalisée est calculée comme suit :

$$C_P^G = \frac{n-1}{\sum_{j=1}^{n} P(i,j), i \neq j} \quad (6.3)$$

pour un nœud i, où P(i, j) est la longueur du chemin le plus court entre les nœuds i et j.

La distance géodésique entre deux nœuds dans un graphe est le nombre d'arêtes dans un plus court chemin qui les relie. En cas des nœuds non connectés, la distance géodésique est mise à '1'.

L'excentricité du nœud est la plus grande distance géodésique qui sépare un nœud i des autres nœuds du réseau. Elle indique donc à quelle distance un nœud i est loin de j. Les nœuds avec une faible excentricité peuvent être utilisés pour diffuser l'information rapidement et jouer le rôle des autoritaires de flux d'information.

En se basant sur ces mesures, des auteurs ont classé les nœuds en deux classes : critiques et non critiques (Sarr *et al.* 2012). Les nœuds critiques sont ceux qui jouent un rôle central dans le réseau avec des scores de centralité élevés. On trouve trois types de nœuds : le leader, l'intermédiaire, et le témoin. Le leader est le nœud qui a une centralité de degré élevée, il interagit avec de nombreuses autres entités. Le médiateur est celui qui agit comme une entité intermédiaire entre les groupes. Il a une centralité d'intermédiarité élevée. Et enfin le témoin qui a la meilleure visibilité sur les flux de l'information dans le réseau, ayant une centralité de proximité élevée. Les nœuds non critiques quant à eux, sont les nœuds avec un rôle moins important. On trouve deux catégories : un *Finger*, c'est le nœud dont la mesure de centralité s'écarte légèrement de celle d'un nœud critique de type leader, intermédiaire ou témoin. Ceci permet d'avoir *k-Fingers*. Deuxième catégorie des nœuds non critiques est le *Follower* qui n'est ni un nœud critique ni un nœud de type *Finger*. En effet l'identification de ces nœuds centraux dans un graphe représente un enjeu important dans plusieurs domaines.

Nous passons maintenant à l'approche proposée qui présente une modélisation d'un réseau social composé d'un ensemble d'utilisateurs. Ces utilisateurs sont regroupés en communautés d'intérêts. Chacune regroupe un sous-ensemble d'utilisateurs qui sont liés les uns aux autres par plusieurs types de relations, et dans lequel les individus les plus importants sont identifiés. Ils permettent d'assurer une meilleure assistance aux nouveaux utilisateurs durant leurs premières utilisations du système, et cela en visant à leur offrir une meilleure exploration du contenu. Avant d'entamer la modélisation de notre réseau social, nous commençons par un petit exemple concret qui permet de donner une vision plus claire sur la problématique étudiée. Il est à noter que notre réseau social n'est pas limité à cet exemple. L'objectif de ce scénario est uniquement de donner une meilleure compréhension du problème et illustrer l'approche globale proposée à travers une étude de cas.

VI.5. Approche proposée

VI.5.1. Scénario illustratif du problème

Considérons l'exemple d'un département de recherche universitaire constitué de plusieurs équipes de recherche scientifique dont chacune représente une communauté liée à un domaine d'intérêt (Chekkai *et al.* 2011). Chaque membre de la communauté peut avoir des intérêts liés à d'autres domaines. Lorsqu'un nouveau doctorant rejoint le département. Il est considéré comme un nouvel élément pour le département et ne peut pas communiquer aussitôt avec ses membres. Ceci ne lui permet pas de recevoir des recommandations de leur part. Il pourra par contre prendre contact avec les représentants des équipes qui ont des profils similaires à son sujet de recherche. Ces derniers vont lui recommander de la documentation qui peut répondre à ses besoins, et/ou lui mettre en contact avec d'autres membres de leurs équipes qui travaillent sur des domaines d'intérêt liés à son sujet de recherche. Ils peuvent aussi lui recommander d'autres membres associés à d'autres domaines afin qu'il puisse enrichir et évoluer ses intérêts. Partant de cet exemple, nous proposons de modéliser cette problématique par une représentation graphique d'un réseau social dont les nœuds représentent les communautés. Chaque communauté est représentée par un individu qui représente le point d'entrée à sa communauté. Cette communauté représente un sous-ensemble d'utilisateurs ayant des caractéristiques commun. Ils sont représentés par un sous-graphe social. Les communautés du réseau peuvent être liées les unes aux autres, car un utilisateur

peut avoir plusieurs centres d'intérêt et peut donc appartenir à plusieurs communautés. Afin de faciliter l'exploration de ce réseau, une carte hiérarchique est proposée. Elle représente un guide d'exploration du contenu de ce réseau social (cf. figure 6.4).

VI.5.2. Modélisation du réseau social

La modélisation de notre réseau social reflétant les utilisateurs réguliers du système s'effectue en trois étapes: la définition du ou des critère(s) de formation des communautés, la découverte des communautés, et la modélisation du graphe social de chaque communauté.

Critères de formation des communautés. La première étape consiste à identifier le ou les critères sur lesquels le système s'appuie pour construire les communautés des utilisateurs. Selon ces critères, la position des utilisateurs et leur regroupement en communautés sont susceptibles de varier. Ces critères dépendent principalement du contenu du profil de l'utilisateur et de sa structure. Ainsi, la définition de ces critères de formation de communautés se base sur les mêmes niveaux de représentation des données d'intérêt des utilisateurs qui sont définis par le système d'analyse QF proposé dans le chapitre 4. Cela permet d'avoir une carte offrant plusieurs critères d'exploration du contenu du réseau social. Le premier niveau propose à l'utilisateur un ensemble de groupes de SFC à explorer, lorsque l'utilisateur fait un choix, les sujets qui sont liés au groupe sélectionné sont proposés, et ainsi de suite jusqu'au niveau élémentaire constitué de domaines d'intérêt.

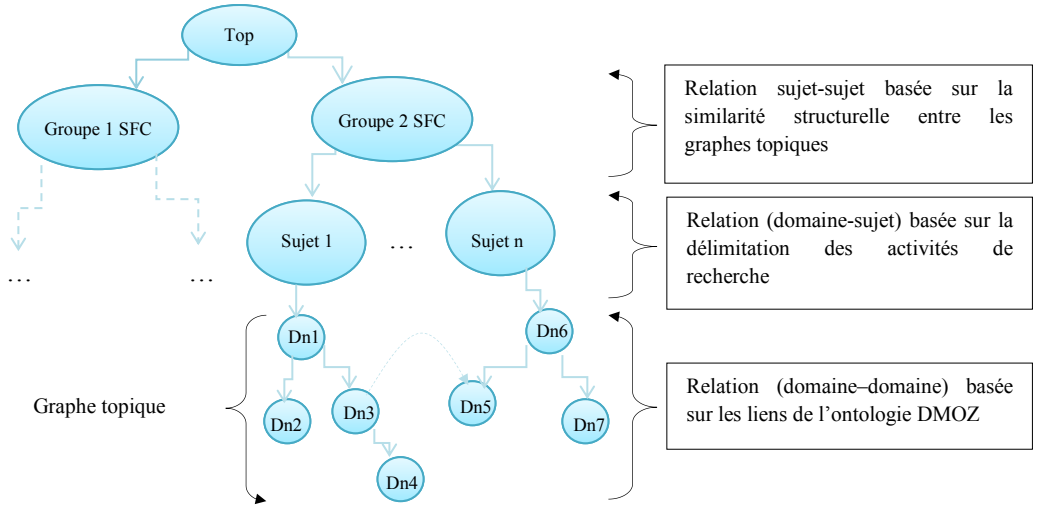


Figure 6. 4. Carte d'exploration du contenu d'un réseau social

Découverte de communautés : Chaque niveau de la carte correspond aux utilisateurs qui sont regroupés selon un critère donné. Par exemple, dans le premier niveau les communautés sont formées à partir des groupes de SFC. Chaque communauté regroupe les utilisateurs ayant un intérêt pour un groupe SFC. Nous avons vu dans la section V.5.4.4.1 comment un utilisateur est affecté à un groupe de SFC. De la même manière, le regroupement des utilisateurs est effectué dans les autres niveaux 2 et 3, respectivement par sujets d'intérêt et par domaines d'intérêt. Un utilisateur peut appartenir à une ou plusieurs communautés, car il peut par exemple être intéressé par un ou plusieurs sujets de recherche. Cependant, les données d'intérêt qui sont associées à un utilisateur u_i au sein d'une communauté C_i sont filtrées selon le critère de son appartenance à cette communauté.

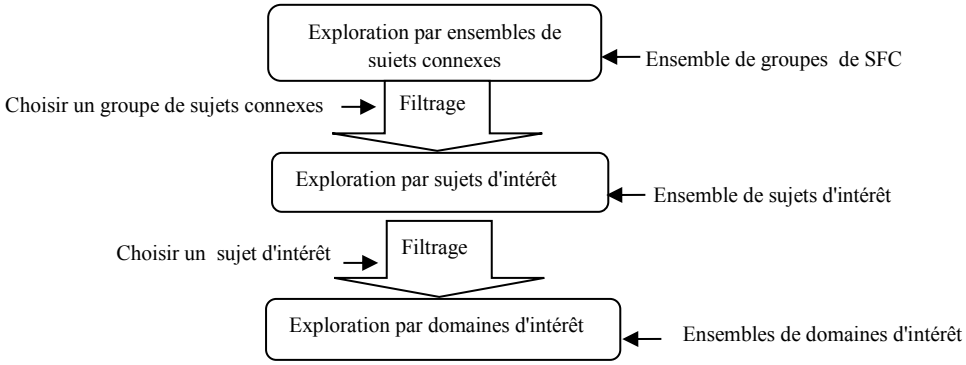


Figure 6. 5. Niveaux d'exploration dans une carte de communautés d'utilisateurs

À chaque fois que l'utilisateur fait un choix dans la carte, la communauté d'utilisateurs qui correspond à son choix est proposée (cf. figure 6.5).

Modélisation du graphe social de chaque communauté: Il consiste à relier les utilisateurs au sein de chaque communauté. Soit $G_1 = (V_1, E_1)$ un sous-graphe social d'un ensemble de chercheurs scientifiques noté par V_1, ils sont connectés les uns aux autres par des relations notées par E_1 (cf. figure 6.6)**.**

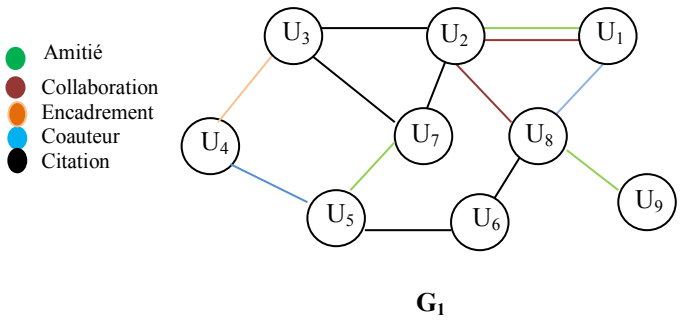


Figure 6. 6. Exemple d'un graphe social d'un groupe de recherche scientifique

Plusieurs éléments sont considérés dans la modélisation d'un graphe social G_i d'une communauté d'intérêt. Ils sont définis comme suit :

Individu. Un individu représente un utilisateur du système répondant au critère de formation d'une communauté. Dans notre exemple, cet utilisateur peut être soit un enseignant, un étudiant, un ingénieur de recherche ou autre. Chaque utilisateur est représenté par un profil qui décrit ses centres d'intérêt.

Relation. Une relation est le lien qui existe entre deux utilisateurs dans la communauté. Le type d'une relation diffère selon le contexte du système. Cette relation peut-être de nature professionnelle, telle que l'encadrement, la collaboration de travail, une relation de citation dans une publication scientifique, relation de coauteur, partage de données, etc., ou de nature personnelle telle que l'amitié, l'échange de messages, etc., ou elle peut être sémantique qui représente une corrélation entre les profils d'intérêts de deux utilisateurs. Compte tenu du fait que certains utilisateurs interagissent avec leurs utilisateurs voisins par l'échange de ressources plus qu'ils consomment des ressources sur le système, il sera utile de considérer de telles relations sociales dans la représentation du graphe social. Ainsi, les interactions entre deux utilisateurs sont également prises en compte pour définir une relation entre ces deux individus.
La signification de tous ces types de relation reste la même, ils définissent pour nous une proximité entre deux individus. La détection de la similarité sémantique permet de relier deux individus qui n'ont jamais interagi ensemble. Ces individus peuvent appartenir à la même communauté ou à des communautés différentes. Cette relation sémantique peut être fructueuse lors de l'exploration de données.

La distinction entre les liens forts et faibles a eu un fort impact sur le développement de l'analyse des réseaux sociaux (Granovetter 1973). Une relation peut donc être considérée comme superficielle (ex. amitié sans aucun échange de ressources), ou comme une relation forte reliant deux individus en

collaboration, codirection, une relation sémantique, échange de ressources, etc. Nous proposons à cet effet une hiérarchie de relations illustrée dans la figure 6.7.

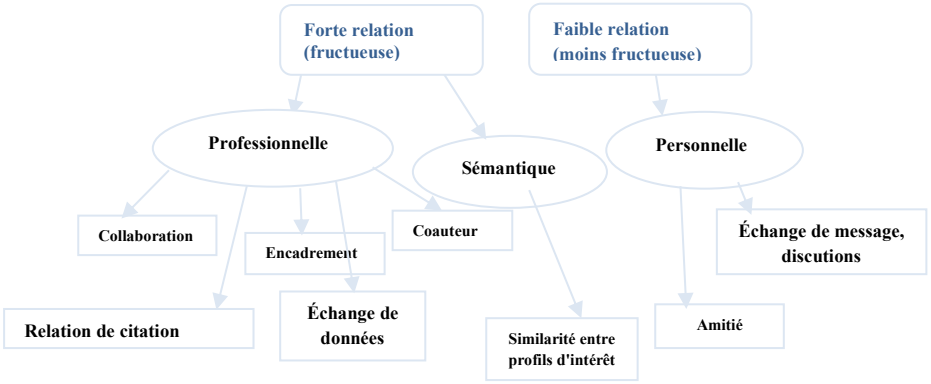


Figure 6. 7. Hiérarchie de relations possibles entre les utilisateurs système

Une relation peut être simple, c'est à dire, une seule relation à la fois qui relie un couple d'individus. Elle peut être simultanée lorsque plusieurs types de relations existent en même temps. Ainsi, nous proposons de distinguer entre les types de relations qui relient les mêmes couples d'individus. Cela prévient la déconnexion entre les individus dans le graphe social lorsqu'une des relations disparaît (par choix de l'utilisateur de du système). Cette disparition peut nuire à l'exploration de données au sein du réseau social. La distinction des relations se manifeste par exemple dans le graphe G1 entre les deux individus U_1 et U_2 à travers deux relations différentes (ex. amitié et collaboration). Cela empêchera leur déconnexion lorsque par exemple l'engagement de collaboration prend fin entre ces deux utilisateurs. Si cela se produit, les deux individus sont toujours considérés comme des voisins dans le graphe à travers une relation d'amitié et l'exploration de données est toujours possible entre eux. Afin de minimiser le nombre d'arêtes dans le réseau, les relations reliant le même couple d'individus et appartenant au même contexte d'interaction (cf. définition ci-après) sont fusionnées en un seul lien qui représente une relation générique.

Contexte d'interaction : deux relations sont considérées comme étant liées au même contexte d'interaction si elles appartiennent à la même hiérarchie de relations (cf. figure 6.7).

Pondération d'une relation. Nous avons choisi de limiter le poids d'une relation simple à deux valeurs 0 et 1, où la valeur 1 correspond à une relation très forte entre deux utilisateurs (relation fructueuse) et la valeur 0 correspond à une relation moins intéressante. Cette pondération représente donc le degré

d'importance de la relation. Une relation simultanée quant à elle a un niveau de pondération égal à la somme des pondérations des diverses relations simples qui la constituent. Ce score est appelé dans notre modèle par le score de multiplicité.

Évolution du réseau. Le réseau évolue au fur et à mesure que les différentes interactions des individus progressent. Une relation simple entre deux individus peut progresser. Par exemple, deux individus travaillant ensemble peuvent devenir amis et cela se traduit par l'ajout d'un lien entre leurs nœuds représentatifs dans le graphe social. En outre, si les deux relations sont dans le même contexte d'interaction, aucun lien n'est ajouté entre ces individus. Le score de pondération est simplement augmenté de 1. Une relation simultanée peut aussi devenir simple lorsque deux individus sont connectés par deux relations dans le même contexte et l'une d'elles prend fin (ex. une collaboration et une relation d'encadrement). La disparition de cette relation est modélisée par une diminution de son degré de multiplicité.

Type du graphe social. Nous considérons les interactions entre les individus comme des relations fructueuses dans les deux directions. Nous parlons donc de relations non orientées et d'un graphe non orienté. Chaque lien est pondéré avec un degré d'importance dans le cas d'une relation simple ou par un degré de multiplicité en relation simultanée. Le graphe social est dans ce cas dit graphe pondéré.

Profil de la communauté. À chaque communauté est associé un profil. Il s'agit d'une vue globale sur le contenu proposé par cette communauté qui peut être obtenu à travers un survol de la souris sur le nœud qui correspond à cette communauté. Ce profil est défini comme étant un nuage d'éléments (domaines d'intérêt fréquents, requêtes et d'étiquettes fréquentes) qui apparaissent le plus dans les profils des utilisateurs de la communauté cible. Ces éléments sont mis en avant durant la tâche d'exploration pour donner à l'utilisateur une vue globale sur le contenu de la communauté qu'il s'apprête à explorer.

VI.5.3. Connectivité entre utilisateurs basée sur la qualité du flux d'information

L'évaluation de la similarité sémantique entre les utilisateurs permet d'un côté de relier, dans le graphe social, les individus qui n'ont jamais interagi ensemble, et d'un autre, de mettre à jour les pondérations existantes avec des scores de similarité. Si le nombre de liens sémantiques à considérer dans le graphe est élevé, cela peut être problématique. Il peut prolonger la durée d'exploration de données (Sarr *et al.* 2012). Pour cela, nous définissons une solution qui contrôle la croissance de la durée de propagation

d'information. Cette solution consiste à relier uniquement les individus qui présentent une forte similarité. Il consiste à ajouter les liens de manière itérative et dans l'ordre décroissant tout en maintenant une bonne qualité de propagation d'information. À cet effet, nous introduisant un seuil de similarité Ω, permettant de définir si un lien l_i (u_1,u_2) entre deux utilisateurs u_1 et u_2 est suffisamment fort pour être retenu, et proposons un algorithme d'enrichissement de liens à base de qualité de flux d'information dans un réseau social.

Nous nous inspirons des travaux de Idrissa et ses collègues pour déterminer le nombre optimal de liens à considérer dans l'enrichissement du graphe social de telle sorte que la diffusion de l'information ne diminue pas beaucoup par rapport à l'état initial (Sarr *et al.* 2012). La qualité de diffusion est déterminée en fonction de l'excentricité du nœud témoin. Celui-ci représente le nœud le plus proche de n'importe quel autre nœud dans le réseau et son excentricité détermine le temps qu'il faut à une information de parvenir à n'importe quel autre nœud dans le graphe (cf. algorithme 6.1). Cela permet d'estimer dans notre cas le temps d'exploration du contenu dans le réseau social.

Entrée :
G = (V, E): le graphe social initial,
$L_S = \{l_1,.....l_N\}$: liste des liens sémantiques à ajouter,
α: écart de degré de propagation de l'information,
β : seuil de similarité

Sortie :
G' : le nouveau graphe social

Début
N_S ← Identifier le nœud témoin dans S
$e(N_S)$ ← Déterminer l'excentricité de Ns

Pour i de 1 à N **faire**
Si $L_S(i) \in E$ **alors** // Si le lien existe déjà dans le réseau

Pondérer $L_S(i)$ dans S,

Sinon Si (Sim (l_i)- Ω > 0) **alors** // Si le lien a une similarité forte
G' ← Ajouter $L_S(i)$ dans S
$N_{S'}$ ← Identifier le nouveau nœud témoin dans G'
$e'(N_S)$ ← Déterminer l'excentricité de $N_{S'}$
Si $(e'(Ns) \leq \alpha \times e(Ns)+e(Ns))$ **alors**
i++;
Sinon
retirer (G', Ls(i));
i++
Fin pour
Retourner S'
Fin

Algorithme 6. 1. Algorithme de connexion sémantique entre utilisateurs

VI.5.4. Identification des utilisateurs importants dans une communauté

Tel qu'il a été abordé précédemment, les mesures de centralité permettent d'identifier le rôle et la position des nœuds dans le graphe. Notre objectif est de sélectionner les utilisateurs ayant un grand nombre de contacts et avec lesquels l'exploration du contenu s'effectue avec un minimum d'efforts. Ainsi sur la base des mesures présentées dans la section VI.4.2 nous avons éliminé la centralité du degré qui définit le nombre de membres avec lesquels un nœud est lié directement. Celui-ci ne répond pas à nos besoins, car il prend en considération uniquement le voisinage immédiat du nœud. Le degré d'intermédiarité est choisi, il exprime le degré de connectivité d'un utilisateur avec les autres membres du graphe. Le degré de proximité est également choisi, il détermine le degré avec lequel l'utilisateur est près de tous les autres utilisateurs. Ceci permet d'avoir une accessibilité rapide vers les autres membres.

Puisque nous avons affaire à un graphe pondéré, nous tenons en compte de ces pondérations qui représentent la solidité des liens entre les utilisateurs. Ces liens sont considérés dans notre cas comme un degré de profit entre les utilisateurs et permettent de déterminer si un chemin peut être utile ou non pour une exploration pertinente. Ainsi, les individus importants sont également sélectionnés vis-à-vis à ce critère d'exploration. Il s'agit des utilisateurs qui possèdent aussi le chemin le plus fructueux pour y accéder à leurs contacts. Nous appelons ce score par le « score de profit ». Ainsi, un score de profit d'un individu « i » est égal à la moyenne des scores de pondérations de ses chemins avec ses contacts «Co».

$$\mathrm{Score_P}(\mathrm{i},\mathrm{Co}) = \frac{\sum_{\mathrm{j}=1}^{\mathrm{N}} \mathrm{Score}(\mathrm{i},\mathrm{j})\mathrm{i} \neq \mathrm{j}}{\mathrm{N}} \qquad (6.\ 4)$$

Où N est le nombre des contacts de l'individu « i » dans sa communauté. *Score (i, j)* est le score de pondération d'un chemin reliant deux individus i et j , il est égal à la moyenne des degrés de pondération dans un plus court chemin H qui relient ces deux individus.

$$\mathrm{Score}\ (\mathrm{i},\mathrm{j}) = \frac{\mathrm{P(H)}}{\mathrm{K}} \qquad (6.\ 5)$$

Où K est le nombre d'arcs dans le chemin H, et P(H) est la longueur de H (cf. la définition ci-dessous). Il est égal à la somme des pondérations qui sont associées à ses arcs.

Chemin. Un chemin dans un graphe est une suite finie d'arcs consécutifs.

VI.5.4.1. Mesure d'importance composée

Nous introduisons une mesure d'importance composée pour évaluer l'efficacité d'un individu au sein de sa communauté. La mesure est définie comme suit :

$$IC_i = (C_P, C_I, Score_P) \qquad (6.\ 6)$$

Où C_P est le degré de proximité, C_I est le degré d'intermédiarité, et $Score_P$ est le score de profit.

Revenons à l'exemple de la figure 6.5 sur lequel nous appliquons ces mesures. Le tableau 6.1 illustre les valeurs de proximité, d'intermédiarité des nœuds qui sont calculées par le logiciel d'UCINET, et le score de profit calculé à travers l'équation 6.4. La sélection d'un meilleur utilisateur revient à effectuer une comparaison multidimensionnelle entre les utilisateurs en utilisant la notion de dominance. Le résultat représente l'utilisateur optimal au sens de Pareto (cf. définition 3.28), il correspond à un meilleur compromis entre les valeurs de centralités considérées dans le tableau 6.1.

Utilisateur	**score1: centralité d'intermédiarité**	**score2: centralité de proximité**
U_1	0	0.424
U_2	0.371	0.482
U_3	0.322	0.518
U_4	0.067	0.466
U_5	0.439	0.482
U_6	0.06	0.368
U_7	0	0.285
U_8	0.036	0.437
U_9	0	0.466

N	**score3: degré de profit**
U_3	0.375
U_5	0.2

⇨ Responsable = U_3

Tableau 6. 1. Valeurs de centralités et degré de profit des utilisateurs de la communauté

En se référant aux résultats présentés dans le tableau 6.1, nous pouvons voir que le Skyline résultant est l'ensemble d'individus U_3 et U_5. Pour sélectionner l'individu responsable parmi les deux individus résultants, nous nous basons sur le degré de profit. Ainsi l'utilisateur U_3 est le responsable de sa communauté.

VI.5.5. Construction du profil d'un nouvel utilisateur

L'analyse comportementale de l'utilisateur en fonction de ses des interactions à travers la carte des communautés permet au système d'affiner ses groupes d'intérêt. Ces derniers peuvent être utiles pour alimenter le système d'inférence d'intérêts qui se base sur l'exploitation de règles d'association structurées elles-mêmes en groupes d'intérêt (cf. section V.5.4.4). Cette inférence contribue à la construction et à l'enrichissement de son profil.

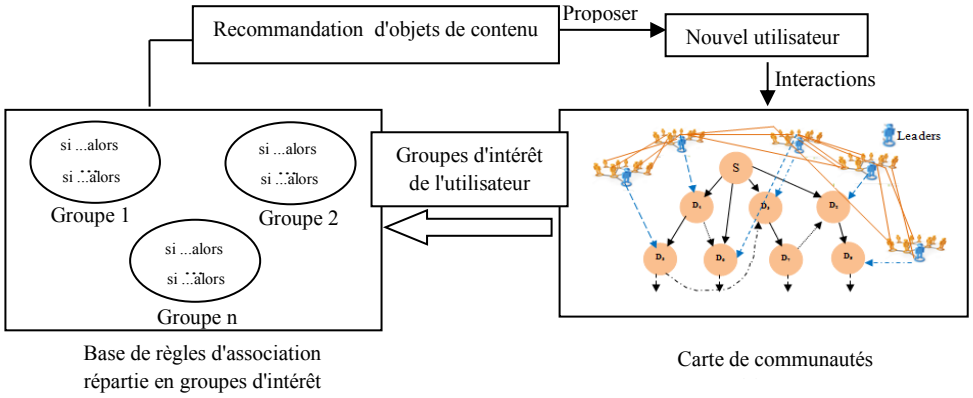


Figure 6. 8. Construction du profil d'un nouvel utilisateur

Les règles d'association sont exploitées par le système comme une base de connaissance qui aide à enrichir les intérêts d'un nouvel utilisateur en inférant d'autres intérêts correspondant à ses interactions qui sont effectuées sur le système à travers la carte proposée. Nous appelons ces interactions par les éléments d'intérêt de l'utilisateur, ils peuvent être de différentes catégories selon le niveau d'exploration : sujets d'intérêt, domaines d'intérêt, requêtes, étiquettes, documents, etc. Ils représentent les faits qui alimentent le processus d'inférence. Ce processus se base sur les étapes suivantes :

- Le système calcule le groupe d'intérêt de l'utilisateur qui permet de cibler les règles d'association dans la base de connaissance qui peuvent être exploitées dans l'inférence d'intérêts. Cela est fait en calculant la proportion d'éléments d'intérêt de l'utilisateur qui appartiennent à chaque groupe de ses interactions par rapport à la proportion de tous ses éléments d'intérêt durant sa session de recherche. Le groupe d'intérêt ayant une grande proportion est utilisé pour sélectionner un ensemble de règles (cf. figure 6.8).
- Le système vérifie ensuite dans l'ensemble des règles sélectionnées s'il existe des règles ayant comme antécédents les éléments auxquels l'utilisateur s'est intéressé et lui propose les documents qui sont liés

aux objets conséquents de ces règles. L'utilisateur interagit avec la liste des documents pour valider ceux qui sont intéressants pour lui. Ce processus est résumé dans la figure 6.9.

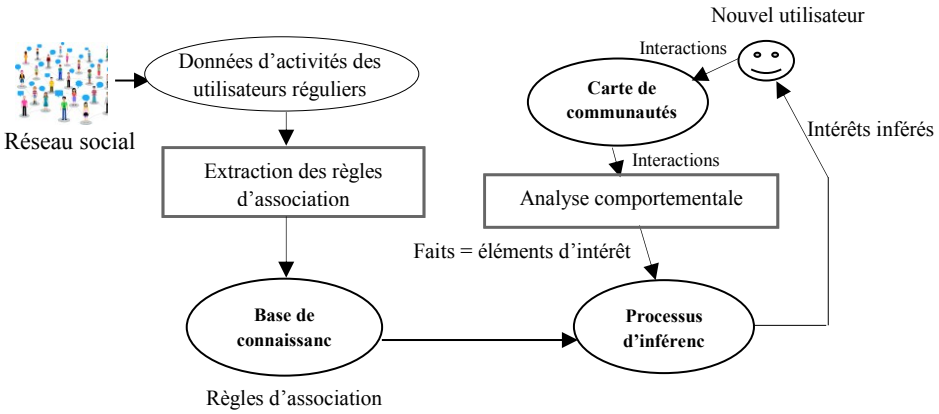


Figure 6. 9. Processus d'inférence d'intérêts pour un nouvel utilisateur

VI.6. Conclusion

Nous avons présenté dans ce chapitre une nouvelle approche pour pallier le problème de démarrage froid d'un nouvel utilisateur dans un système de recommandation. Cela passe par l'utilisation d'une carte de communautés d'utilisateurs. Cette carte organise le système en plusieurs communautés qui sont formées selon différents critères. L'identification des utilisateurs centraux au sein de ces communautés offre la possibilité d'avoir une exploration facile et fructueuse du contenu. Cette mesure d'importance se base sur plusieurs critères inspirés du domaine de l'analyse des réseaux sociaux. Cette solution peut être utile pour extraire les intérêts d'un nouvel utilisateur et construire son profil.

Chapitre 7 : Protocole d'implémentation et d'évaluation d'un système de recherche d'information multi-facettes

VII.1. Introduction

Nous avons présenté dans les chapitres précédents des modèles théoriques visant à tirer parti des deux dimensions sémantique et sociale du web pour améliorer le processus de RI. Nous avons aussi tiré profit du concept de la recherche par facettes pour la proposition d'un modèle d'indexation et de recherche multidimensionnelle. Également, nous avons profité de différents domaines tels que le domaine de l'analyse et l'extraction de données ainsi que celui des théories des graphes pour la représentation et l'enrichissement du modèle utilisateur qui traduit ses centres d'intérêt et ses préférences. Ce dernier est exploité pour personnaliser les recherches d'information de l'utilisateur au sein d'un modèle de recherche centrée-utilisateur en vue d'améliorer davantage la pertinence des résultats de recherche. Le chapitre présent décrit les détails de la mise en œuvre de ces modèles afin d'évaluer et de démontrer leur efficacité au sein d'un SRI qui se base sur les facettes de données (SRIF). Pour cela, l'implémentation d'un prototype fonctionnel est effectuée, et un cadre d'évaluation est nécessaire.

Nous avons vu, dans le chapitre de l'état de l'art (section II.1.5 et II.2.4), comment les SRIs peuvent être évalués. Nous avons parlé du premier modèle Cranfield (Cleverdon 1967) adopté pour l'évaluation des SRIs classiques et ses limitations en présence de la dimension utilisateur. Cette évaluation classique est fondée sur l'utilisation d'une collection de test où i) les requêtes sont les seules ressources qui représentent le besoin informationnel de l'utilisateur, et ii) la pertinence des documents de tests est purement thématique, c'est-à-dire, elle est estimée à travers une similarité fondée sur le critère de sujet. En outre, cette pertinence est considérée par des experts qui ne peuvent pas toujours se mettre à la place des utilisateurs et estimer pertinemment leurs attentes, d'où l'émergence des approches d'évaluation orientée-contexte qui sont adaptées à l'évaluation des SRIP. Ces approches mettent en place des scénarios d'évaluation qui impliquent l'utilisateur dans le processus d'évaluation. Cela consiste à l'intégration du profil utilisateur comme étant une composante principale de la collection de tests en plus des autres éléments classiques du modèle Cranfield (les requêtes, les documents et les jugements de pertinence).

Cette intégration du profil permet la contextualisation des jugements de pertinence. Autrement dit, l'évaluation du système est fondée sur une pertinence telle qu'elle est vue par l'utilisateur, dite pertinence perceptionnelle ou situationnelle (cf. section II.1.2.1). Par conséquent, une liste de jugements de pertinence doit être différente pour chaque requête de chaque utilisateur.

Dans un cadre social, la collection de tests utilisée pour évaluer les systèmes sociaux se distingue par i) une liste d'étiquettes à la place de la liste des requêtes, et ii) un jugement de pertinence qui se base sur le comportement social de l'utilisateur envers les ressources proposées, à savoir son comportement d'annotation, au lieu de son comportement classique (clics, consultation des pages, évaluations des pages, etc.). Des exemples de telles collections sont les collections Delicious, Flickr, Diigo, etc. qui offrent un ensemble d'étiquettes et des jugements de pertinence centrée utilisateur associant à chaque étiquette de chaque utilisateur les ressources pertinentes.

Peu importe le cadre d'un SRI, son évaluation peut être effectuée soit par simulation d'utilisateurs ou par des utilisateurs réels. Dans la première catégorie, l'évaluation intègre un scénario d'évaluation défini par des interactions hypothétiques avec le SRI permettant de simuler des utilisateurs et construire leurs profils (Mostafa *et al.* 2003). Ainsi, l'évaluation de l'efficacité du système est basée sur l'utilisation des jugements de pertinence qui sont prédéfinis dans une collection de tests. Dans la deuxième catégorie, des utilisateurs réels sont impliqués dans l'évaluation du système. Pour ce faire, des interactions telles que les clics, le temps de consultation des pages, les évaluations, les annotations, etc. sont exploitées à la fois pour l'apprentissage des profils des utilisateurs et dans l'évaluation de l'efficacité du système. En dépit du grand avantage offert par cette deuxième catégorie d'évaluations qui se traduit par le réalisme des besoins en information des utilisateurs, elle est malheureusement coûteuse en temps puisqu'elle implique des utilisateurs réels. En outre, ces derniers ne sont pas toujours disponibles pour cette tâche d'évaluation.

Dans les deux cas (utilisateurs simulés ou réels), l'efficacité est estimée par l'évaluation de l'impact du profil utilisateur lorsqu'il est intégré dans le processus de RI.

Le problème de l'évaluation est d'autant plus particulier dans un SRI multidimensionnel et hybride comme le nôtre. Selon nos connaissances, aucun travail dans la littérature n'a traité la recherche multidimensionnelle dans un contexte hybride où l'interaction de l'utilisateur avec le système ne se limite pas à la soumission des requêtes de recherche et à l'exploration des listes de résultats, mais aussi à l'annotation des documents qui sont proposés par le système et à l'exploration des résultats selon des

facettes de données et d'un ensemble de requêtes prédéfinies sur l'interface multidimensionnelle que nous considérons comme des valeurs de facettes et les appelons par les requêtes système (cf. figure 3.15). Ces valeurs de facettes sont dynamiques et différentes à chaque nouvelle recherche effectuée par un utilisateur en vue d'enrichir son univers de navigation, et cela en se basant sur différents aspects (cf. section 3.6). Elles visent à faciliter l'exploration des résultats et permettent également d'affiner ou élargir la recherche. Le processus d'interaction utilisateur-système a été expliqué dans la section IV.5.1.

Nous constatons que les données de test et le protocole d'évaluation qui sont nécessaires pour l'évaluation de notre système sont particuliers. Cette évaluation nécessite des données qui ne peuvent pas être retrouvées au sein d'une seule collection de tests et certaines d'entre elles ne sont pas disponibles. A nos jours, il n'existe pas un cadre standard et complet qui soit destiné à l'évaluation de tels systèmes. Nous nous sommes intéressés à proposer un nouveau cadre d'évaluation adéquat à notre système tout en s'inspirant de ce qui a été proposé dans la littérature. Ce cadre s'appuie sur la détermination d'un ensemble de tâches d'évaluations. Chacune est destinée à évaluer un modèle qui nécessite une collection de tests spécifique, des métriques d'évaluation, et un protocole d'évaluation adéquat. Ainsi, les taches d'évaluations qui constituent le cadre proposé s'énumèrent comme suit :

- **Tache 1** : Évaluation qualitative de l'efficacité du système multi-facettes (SRIF).
- **Tache 2** : Évaluation de l'efficacité du profil utilisateur multi-niveaux et son modèle d'enrichissement qui se base sur la recommandation de données par inférence collaborative d'intérêts.
- **Tache 3.** Évaluation de l'efficacité des deux modèles de personnalisation de données qui se basent sur l'intégration du profil utilisateur respectivement au niveau de l'indexation des documents et dans le réordonnancement des résultats de recherche.

Pour l'accomplissement de ces tâches d'évaluations, un SRIF est conçu. Cette conception permet d'atteindre un double objectif. Le premier est de construire une collection de tests globale qui répond aux besoins de chacune des taches d'évaluations définies. Le deuxième est d'impliquer de vrais utilisateurs à évaluer le système. Pour la conception de ce système, un prototype fonctionnel d'un SRI est nécessaire que nous commençons par détailler sa mise en œuvre dans la section suivante (cf. section VII.2).

VII.2. Mise en œuvre d'un prototype fonctionnel d'un système de recherche d'information par facettes

Le prototype fonctionnel est développé en utilisant les technologies JEE. Il est construit autour de ElasticSearch (connu aussi sous le nom de « ES » et récemment renommé Elastic). Elastic est un moteur de recherche libre et évolutif basé sur la bibliothèque Lucene (Gormley et Tong 2015), il est caractérisé par ses capacités de répartition de charge et de distribution de données. Par exemple, si plusieurs nouveaux documents sont ajoutés simultanément, ils peuvent être répartis sur plusieurs fragments ou nœuds différents. Il permet aussi d'exécuter plusieurs requêtes de recherche à travers une API multi-recherche. Cette API de recherche prend en charge plusieurs types de requêtes pouvant être exploités et personnalisés selon le besoin.

Le prototype proposé comprend les trois modules principaux trouvés dans les moteurs de recherche traditionnels, à savoir l'exploration de données, l'indexation, et la recherche.

VII.2.1. Module d'extraction et préparation de données

Nous avons développé un robot d'exploration web (appelé en anglais the *Web Crawler*) qui télécharge des pages Web et leurs annotations en se basant sur des liens trouvés dans le site social Delicious. Ce site web offre une diversité de pages web qui sont faciles à analyser. Il est l'un des sites d'étiquetage le plus populaire. Nous avons collecté environ 100 mille documents web annotés d'un ensemble d'étiquettes qui ont été employées par différents utilisateurs pour décrire ces documents. Le contenu textuel de ces documents est extrait à l'aide de l'analyseur de contenu Apache Tika (Tika 2004). Plusieurs métadonnées sont extraites de ce document, tel que le titre, le contenu textuel, l'URL du document, et les mots clés du document s'il existe.

Cet explorateur de données permet d'initialiser une collecte de données par un ou plusieurs mots clés qui peuvent être choisis manuellement. Cela permet de cibler une ou plusieurs thématiques selon le besoin. Les données qui sont recueillies de cette étape représentent la collection de tests de départ que nous exploitons pour étendre son contenu selon le besoin de nos évaluations. Les caractéristiques de cette collection sont illustrées dans le tableau 7.1 ci-dessous.

Entité	Nombre
Documents	112616
Utilisateurs	9420
Étiquettes	43149
$Trels_T$ (jugements de pertinence étiquette-documents)	186400
$Trels_{CU}$ (jugements de pertinence centrée utilisateur)	21841

Tableau 7. 1. La valeur correspondante pour chaque entité dans la collection de données de départ

VII.2.2. Module d'indexation de documents multi-espaces

Le module d'indexation commence par analyser le contenu textuel des documents pour extraire les jetons les plus représentatifs. Ces jetons sont ensuite enrichis à travers plusieurs espaces de représentation qui feront l'objet d'une conception multidimensionnelle de l'index documentaire.

Étape 1: Analyse du contenu. Généralement, un analyseur est constitué de plusieurs composants, chacun effectue un traitement différent. En l'occurrence, nous citons le filtre de caractères, l'analyseur lexical (connu en anglais sous le nom de « *tokenizer* »), et le filtre de jetons. Le filtre de caractères permet de prétraiter le texte avant qu'il ne soit transmis à l'analyseur lexical. Il reçoit le texte d'origine sous forme de flux de caractères qui peut être transformé en lui ajoutant, lui supprimant ou lui modifiant des caractères tels que les caractères spéciaux. L'analyseur lexical découpe le texte filtré en plusieurs jetons en se basant sur différentes techniques de découpage. Le filtre de jetons quant à lui traite les jetons reçus et les transforme selon le besoin (en lemmes ou en racines de mot par exemple).

L'accomplissement de cette étape d'analyse consiste à choisir un analyseur prédéfini ou à construire son analyseur personnalisé à partir de différents composants de base. Dans notre modèle, nous utilisons le module d'analyse proposé par Elastic pour personnaliser cette étape d'analyse. Elle commence par une analyse lexicale standard qui traite un texte d'entrée pour le diviser en mots individuels. Dans cette étape, les mots vides sont conservés pour permettre la recherche d'expressions exactes (titre d'un livre ou film par exemple). Ensuite, les mots composés sont formés en exploitant un découpeur n-Gram appelé en anglais « *the shingle tekonizer* » puis sélectionner les n-Gram ayant une signification sémantique. Cette signification sémantique peut être validée auprès d'un dictionnaire lexical ou une liste prédéfinie de mots composés. Par exemple, le jeton « recette crème glacée » peut être découpé en « recette », « crème », « glacé », « recette crème », « crème glacée », « recette crème glacée ». La taille d'un Gram appartient à

l'intervalle [1-n] et peut être fixée selon le besoin à l'aide des deux paramètres: max_shingle_size et min_shingle_size (Gormley et Tong 2015). Ce type de tokenisation nous permet de conserver les mots composés au lieu de considérer chaque mot séparément. Cela prévient d'affecter le sens de certains jetons qui perdent leur sens lorsqu'ils sont décomposés, en instance nous notons les mots composés tels que « Ice creem », « web site », « remote control », « post office », etc.

Étape 2: Enrichissement et génération de l'index. Cette étape consiste à préparer l'univers de description documentaire puis la génération de l'index multidimensionnel. Cet univers comprend trois espaces. Nous avons l'espace identitaire qui contient les jetons résultant de l'étape de tokenisation. Ce contenu est ensuite enrichi avec des jetons provenant de différentes sources de données externes (sémantique et sociale). Pour préparer l'espace social, nous exploitons les étiquettes d'annotation extraites dans la première étape grâce à l'explorateur de données. Ainsi, à chaque document est associé un ensemble d'étiquettes qui décrivent son contenu. Enfin, nous exploitons le dictionnaire BabelNet pour préparer l'espace sémantique. Ce dictionnaire est riche, il contient des données provenant des dictionnaires WordNet et Wikipédia. Il consiste à interroger BabelNet avec la liste des jetons identitaires et sociaux qui constituent respectivement les deux espaces (identitaire et social) pour extraire les synonymes, hyponymes et hyperonymes de chacun. A chaque document est associé un ensemble de jetons qui le décrivent sémantiquement.

Les différents espaces de description préparés sont utilisés pour la génération de l'index documentaire multidimensionnel. Cette génération se base sur les fonctionnalités proposées par Elastic à travers l'API d'indexation. Nous créons trois espaces d'indexation de trois types différents: l'espace identitaire avec le type « identité », l'espace des étiquettes avec le type « social », et l'espace sémantique de type « ontologie ». L'exemple suivant crée le premier espace identitaire nommé « Space1»:

```
indexRequestBuilder  Space1 = client.prepareindex('contentSpace', 'identity')
```

Chaque document est ajouté aux trois espaces d'indexation et dans chaque espace ce document est indexé avec différents paramètres. Par exemple, dans le premier espace, le document est indexé avec ses jetons identitaires. Ces jetons sont séparés selon deux catégories: les termes et les entités nommées. Une entité nommée est une expression linguistique composée d'un ou plusieurs mots qui identifient un objet à partir d'un ensemble de d'autres objets avec des attributs similaires. Par exemple, citons les personnes, les

entreprises, les lieux géographiques, l'âge, les adresses, les numéros de téléphone, la date, etc. Un terme quant à lui est une unité syntaxique qui peut être composée aussi d'un ou plusieurs mots décrivant un concept dans un domaine. À titre d'exemple, citons un véhicule, une couleur, un fruit, etc. Les termes reflètent alors des notions, tandis que les entités nommées se réfèrent à des objets. Cette séparation nous permet d'utiliser des entités nommées lors de l'interprétation des requêtes (cf. sectionVII.2.4).

Elastic offre une indexation à base du format JSON utilisé pour préparer les documents en champs de description. L'exemple suivant indexe un document dans le premier espace identitaire. Dans cet espace, le document est indexé avec son titre, son contenu textuel séparé en entités nommées et en termes, et avec un chemin url.

```
IndexResponse rep = space1.setSource(jsonBuilder().startObject()
.field('title',title).fied('content',terms).field('contentEn',namedEntities). field('keywords', Listkeywords).field('path', url)
.endObject().excute().actionGet();
```

Cette répartition du contenu en différents champs descriptifs permet de personnaliser la recherche sur un ou plusieurs champs de contenu, ou de privilégier certains d'entre eux par l'utilisation de scores de pondération selon le besoin de l'utilisateur ou selon un paramètre de personnalisation défini par le système. Par exemple, l'utilisateur demande que le titre d'un document ait un poids plus fort dans la recherche que le reste du contenu. Dans notre cas, cette structure est exploitée dans la recherche orientée-utilisateur en vue de personnaliser la recherche sur un sous-ensemble de champs qui décrivent le document selon un utilisateur et son voisinage (cf. section V.2.2.3).

Il est à noter qu'un jeton polysémique relatif à un document donné est désambigüisé avant d'être enrichi et utilisé pour l'indexation de ce document. Une fois les documents sont ajoutés à l'index, ils sont disponibles pour une recherche à texte intégral via une interface utilisateur conviviale ou par une recherche RESTful offerte par l'API REST d'Elastic. La figure 7.1 illustre les étapes constitutives de cette indexation.

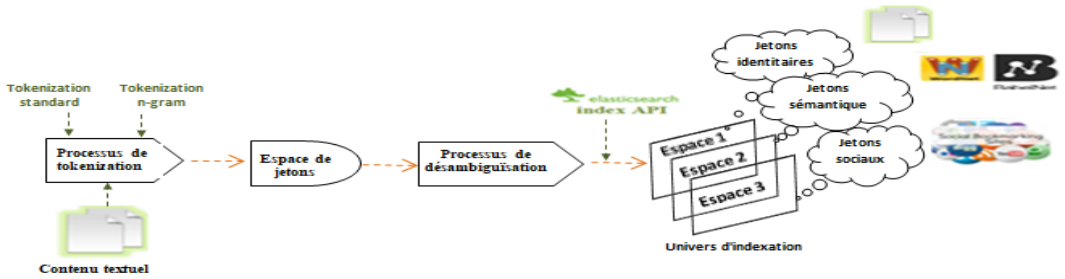


Figure 7. 1. Étapes de la mise en œuvre du processus d'indexation multi-espaces

La section suivante présente comment une requête de recherche est traitée par le système pour associer à son contenu les documents adéquats dans l'index documentaire.

VII.2.3. Module de recherche d'information multidimensionnelle

Le processus de RI est mis en œuvre en plusieurs étapes dédiées principalement à l'interprétation de la requête utilisateur pour interroger l'index documentaire et retourner les documents qui correspondent à cette recherche.

Étape 1: Interprétation de la requête utilisateur. Cette étape consiste en la préparation des dimensions interprétatives du contenu de la requête utilisateur puis l'encapsulation de ce contenu dans une forme exploitable pour l'interrogation de l'index. Tout d'abord, la requête est décomposée en jetons identitaires, cette étape se déroule en plusieurs taches élémentaires: i) l'analyse n-Gram qui est similaire à l'analyse introduite avec le contenu documentaire, puis ii) la sélection parmi les n-Gram résultants les jetons pertinents à la recherche. Cette sélection consiste à i) la reconnaissance des jetons ayant un sens sémantique dans une base de connaissance linguistique. En l'occurrence, nous citons la base de données Wikipedia que nous exploitons à l'aide de l'API JWPL qui offre un accès rapide et efficace à cette base (Ferschke *et al.* 2011). Et à ii) la reconnaissance des jetons composés qui se base sur l'exploitation d'une liste de mots composés recueillie à partir de diverses sources (Merriam 2008) (words)

De la même manière avec le contenu des documents, les jetons résultants de la requête sont divisés en deux catégories de jetons: la catégorie des termes et celle des entités nommées. En général, une entité nommée est exploitée dans une requête pour spécifier par exemple un lieu (pays, ville, adresse), ou une date qui peut être liée à un événement, une personne, ou autre. Par exemple, pour la requête « hôtel

Montréal », le système fait une recherche avec le jeton « hôtel » à travers le champ descriptif des documents « termes » puis filtre parmi les résultats ceux qui sont décrits avec « Montréal » dans le champ descriptif « Entités nommées». Cela permet de considérer « Montréal » comme un filtre et ne pas renvoyer dans de tels cas les pages web qui sont décrites uniquement avec « Montréal ». Pour détecter les entités nommées, nous avons utilisé l'outil Stanford Named Entity Recognizer (NER).

Une fois que les jetons identitaires de la requête sont extraits, le dictionnaire lexical BabelNet est utilisé pour identifier les jetons ambigus (polysémiques). L'ambiguïté de ces jetons est levée à l'aide d'un programme Java que nous avons développé. Ce programme met en œuvre l'approche de similarité composée proposée dans le cadre théorique du chapitre 3 (cf. section III.4.1.2.).

L'enrichissement de la requête est ensuite effectué principalement à base de son contenu original en exploitant des bases de connaissances externes pour relier les jetons identitaires à d'autres jetons. Les dimensions d'enrichissement sont les suivantes :

- **Dimension sémantique fondée sur les relations hiérarchiques entre les jetons dans une base de connaissance lexicale/sémantique** : la préparation de cette dimension s'appuie sur l'exploitation du dictionnaire lexical BabelNet qui permet de relier les jetons identitaires de la requête à leurs synonymes, hyponymes et hyperonymes pour offrir à l'utilisateur des résultats plus spécifiques ou plus génériques.
- **Dimension sémantique fondée sur les sujets de recherche** : en dehors de la similarité hiérarchique qui peut relier deux entités dans une ontologie ou toute autre source de données lexicale, nous définissons la similarité basée sur les sujets de recherche. Cette similarité est plus générale, elle permet de regrouper des requêtes qui sont formulées différemment et sont liées à un même besoin d'information. Cette dimension est préparée en exploitant les clusters sémantiques des requêtes construits au cours des activités de recherche des utilisateurs (cf. Annexe 1). L'idée est donc d'exploiter les requêtes qui couvrent le même sujet de la requête cible. Elles sont recommandées à l'utilisateur comme étant des requêtes prédéfinies sur l'espace sémantique et peuvent l'aider à effectuer d'autres recherches connexes.

Étape 2 : Interrogation de l'index. Cette étape consiste en l'exploitation des espaces d'interprétation obtenus de la requête utilisateur pour associer à leurs jetons les documents correspondants dans l'index multidimensionnel. C'est ce qui est appelé dans le modèle théorique par « la couverture de

jeton ». Pour ce faire, la requête est structurée sous une forme adéquate qui permet de représenter le contenu multidimensionnel préparé. Une requête de type « multi-Query » proposée par l'API de recherche d'Elastic, est alors utilisée. Elle permet d'exécuter plusieurs requêtes de recherche en parallèle, chacune est destinée à représenter une dimension de recherche. Pour chaque dimension de recherche, une requête booléenne est utilisée, elle combine les jetons qui constituent une dimension d'interprétation. Les requêtes booléennes sont des requêtes composées utilisant une ou plusieurs clauses booléennes. Le processus d'interprétation de la requête utilisateur est illustré par la figure 7.2.

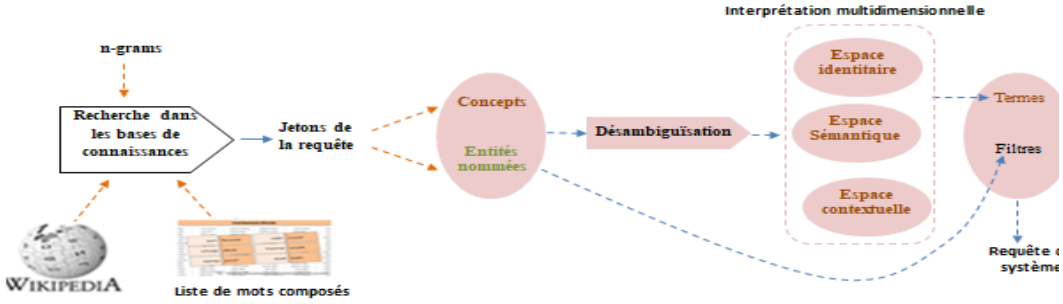


Figure 7. 2. Interprétation multidimensionnelle de la requête de recherche utilisateur

VII.2.4. Interface de recherche et de navigation par facettes de données

L'interface de recherche est le point d'entrée qui donne accès au contenu des ressources du système. L'interface proposée par notre système offre à l'utilisateur la possibilité de créer un compte pour effectuer ses recherches (cf. figure 7.3). Cela permet au système de tracer les activités de recherche relatives aux différents utilisateurs et créer leurs profils. Une fois l'utilisateur est sur son compte, il peut effectuer des recherches en soumettant des requêtes textuelles à travers un champ de texte. Tel qu'il a été expliqué dans le chapitre 3 et présenté dans la figure 3.15, les résultats d'une recherche sont renvoyés sur trois espaces de navigation qui représentent les facettes de données du système. La première facette identitaire représente les documents provenant de l'espace identitaire et ayant correspondu avec le contenu original de la requête de recherche. La deuxième facette sociale est la liste des documents dont les annotations correspondent à un ou plusieurs jetons de la requête utilisateur. La troisième facette sémantique quant à elle représente à la fois les documents qui correspondent au contenu sémantique de la requête dans

l'espace identitaire et social, et ceux qui correspondent à la requête originale dans l'espace d'indexation sémantique.

L'espace de navigation est enrichi avec des valeurs de facettes que le système prédéfinit comme suit :

- **Valeurs de facette basées sur la cooccurrence de jetons identitaires** : tel que le modèle théorique l'a expliqué dans le chapitre 3, il s'agit des jetons qui apparaissent fréquemment et simultanément avec un ou plusieurs jetons de la requête dans le contenu des documents qui résultent de l'espace identitaire notés par D_q. Cette notion de cooccurrence est implémentée comme suit : chaque document est représenté par un ensemble de jetons les plus représentatifs de son contenu. La représentativité d'un jeton est basée sur son importance dans l'index, elle est estimée à travers la pondération BM25F (cf. équation 5.1). Cet ensemble est considéré comme la description qui identifie son contenu, tel que :

  Description (id-doc) :$jeton_1, jeton_2,\ldots, jeton_n$

  À partir de cet ensemble de documents, les ensembles de jetons (appelés les itemsets de jetons) les plus fréquents sont extraits par l'exploitation de l'algorithme Apriori dédié pour de tels calculs. Ce processus d'extraction suit le même mécanisme expliqué dans la section II.2.3.1.5. L'ensemble de jetons ayant une forte fréquence est sélectionné pour représenter les valeurs de la facette identitaire. Dans cette étape, le seuil de fréquence est défini expérimentalement en se basant sur la métrique du support minimum (cf. définition II.1). Pour définir la valeur optimale de cette fréquence, une étape d'apprentissage est nécessaire. Nous faisons varier le support minimum de 0.1 à 1, et à chaque valeur, les itemsets qui répondent au seuil de fréquence sont utilisés pour interroger l'espace identitaire du système. Les itemsets sélectionnés sont ceux qui permettent de localiser le plus de documents parmi la liste de documents de départ qui a été exploitée pour générer ces jetons. Autrement dit, la valeur optimale est celle qui permet d'avoir un meilleur rappel du système. Cela permet de valider à la fois la fréquence des jetons dans les documents de cet espace identitaire et leur représentativité par rapport au contenu de ces documents.

  Cette étape d'apprentissage a été effectuée sur plusieurs requêtes de test Q. Pour chaque requête q, le rappel du système $Rappel_q$ est évalué, puis un rappel moyen est calculé pour toutes les requêtes test.

$$Rappel_{q_i} = \frac{Nombre\ de\ documents\ \ localisés\ parmis\ la\ liste\ de\ départ\ D_q}{Nombre\ de\ documents\ dans\ D_q}$$

$$Rappel\_moyen_Q = \sum_{i=1}^{|Q|} rappel_{q_i}$$

Support-min (seuil de fréquence)	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
Rappel _moyen	0.07	0.12	0.30	0.41	0.28	0.23	-	-	-	-
Nombre moyen de jetons fréquents	84	31	19	11	7	4	0	0	0	0

Tableau 7. 2. Valeurs du rappel moyen du système et du nombre de jetons fréquents obtenus depuis les documents résultants selon différentes valeurs du support minimum

Comme nous pouvons le voir à travers le tableau 7.2, la valeur de 0.4 a permis de donner le meilleur rappel au système. Cette valeur est donc retenue pour extraire les itemsets fréquents qui représentent les valeurs de la facette identitaire.

- **Valeurs de facette sociales du système**: chaque document dans la liste des résultats sociaux (c'est-à-dire, les résultats renvoyés au sein de la facette sociale) est renvoyé avec la liste des étiquettes qui lui sont associées de manière fréquente par les utilisateurs. Cette fréquence est évaluée à travers le nombre d'occurrences de l'étiquette dans l'index documentaire. Cette évaluation est possible grâce à l'API d'agrégations d'Elastic. Les étiquettes sont classées par ordre décroissant de fréquence puis uniquement les dix premières étiquettes sont retenues afin de ne pas saturer l'espace d'affichage. Ces étiquettes représentent les valeurs de la facette sociale que l'utilisateur peut exploiter pour explorer d'autres résultats connexes.
- **Valeurs de facette sémantiques du système :** les instances sémantiques de la requête utilisateur qui sont extraites à base de sujet de recherche sont offertes sous forme d'une liste de recommandations dans la facette sémantique. Ces requêtes représentent les valeurs de la facette sémantique.

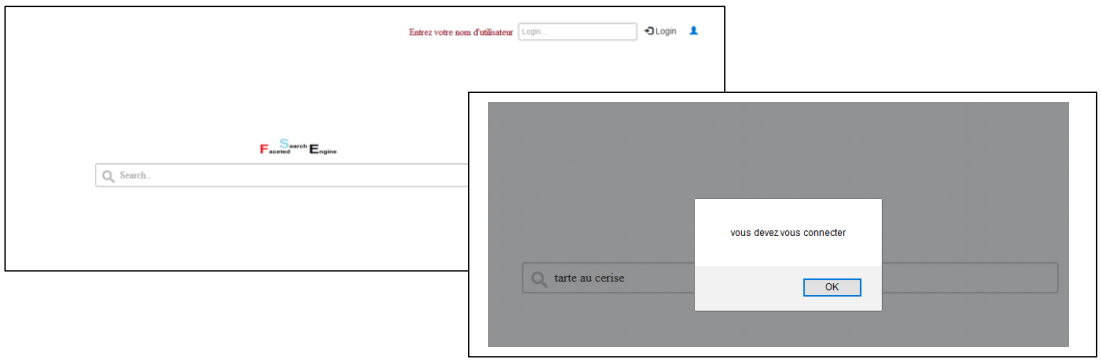


Figure 7. 3. Interface de recherche utilisateur

Nous avons montré dans cette section comment notre SRI est conçu. La section suivante présente le cadre d'évaluation de ce système.

VII.3. Cadre d'évaluation d'un SRI multidimensionnel

Dans cette section, un cadre d'évaluation de l'efficacité du système est proposé. Il s'appuie sur la réutilisation de la collection de données de départ illustrée dans le tableau 7.1 en vue de l'adapter aux besoins du système. Cela est effectué en augmentant le contenu de cette collection avec d'autres composantes qui sont prises en compte dans la construction et l'évaluation de nos modèles. Tel qu'il a été vu dans la section précédente, il consiste à i) indexer l'ensemble des documents de cette collection par un système fonctionnel, ii) donner aux utilisateurs l'accès aux documents indexés à travers une interface de recherche multi-facettes, et finalement iii) analyser les interactions de ces utilisateurs avec le système en vue de collecter les données d'intérêt nécessaires. Ces données d'intérêt représentent les profils de ces utilisateurs qui contribuent à une évaluation centrée-utilisateur du système.

Le cadre d'évaluation proposé se base sur l'utilisation des composantes suivantes que nous détaillons ci-après étape par étape leur construction :

- Une collection de tests étendue qui comprend :
    - Une collection de requêtes soumises par des utilisateurs réels. Chaque requête est annotée d'un sujet d'intérêt et d'un ensemble de domaines d'intérêt. Les requêtes sont collectées via

le système conçu, et les annotations (notamment les sujets et les domaines) sont obtenues via un questionnaire.

- Une collection de documents à interroger fournie par la collection de départ du tableau 7.1.
- Une collection d'étiquettes fournie par la collection de départ.
- Les profils utilisateurs construits suite à des interactions réelles des utilisateurs, ils englobent :
  - Les centres d'intérêt de chaque utilisateur où chacun représente un sujet d'intérêt et englobe toutes les données d'activités (documents, requêtes et étiquettes) relatives à un même besoin d'information.
  - Les sujets fortement connexes qui représentent des clusters contextuels des données d'activités de l'utilisateur.
  - Les jugements de pertinence qui associent pour chaque utilisateur i) les documents pertinents aux requêtes de recherche, ii) les documents pertinents aux étiquettes utilisées.
  - Les jugements de pertinence par facette d'intérêt. Ils sont notés par $Qrels_{u,fe_i}$qui associent pour chaque facette de données fe_i les documents pertinents aux requêtes de l'utilisateur u.

- Une stratégie d'évaluation pour chacune des tâches d'évaluations définies dans ce cadre d'évaluation.

VII.3.1 Construction d'une collection de tests étendue

Le principe de construction de cette collection de tests est basé sur les hypothèses suivantes :

- **Hypothèse 1-Pertinence perceptionnelle/situationnelle.** La pertinence est estimée selon la perception de l'utilisateur.
- **Hypothèse 2-Pertinence graduée et évolutive.** Les données pertinentes d'un utilisateur n'ont pas le même degré de pertinence, d'où une pertinence graduée de ces données qui se base sur l'aspect fréquentiel. Cette pertinence n'est pas statique, elle est évolutive au fil du temps.
- **Hypothèse 3-Pertinence individuelle.** Chaque requête/étiquette est considérée comme étant différente et représentative d'un besoin d'information différent quand elle est émise par différents

utilisateurs d'où l'existence de différents ensembles de jugements de pertinence pour cette requête/étiquette.

- **Hypothèse 4-Pertinence contextuelle.** Deux requêtes/étiquettes lexicalement similaires peuvent être liées à plusieurs interprétations. Elles peuvent donc représenter différents besoins chez le même utilisateur lorsqu'elles sont utilisées dans différents instants. Pour une exploitation pertinente de ces données, celles-ci doivent être annotées d'un contexte qui désambiguïse son contenu.
- **Hypothèse 5. Documents pertinents.** La liste de tous les documents pertinents est connue.

Pour la construction de notre collection de tests, nous avons comme données de départ un corpus de documents accessible à travers le SRIF conçu, et un ensemble d'utilisateurs réels (étudiants universitaires) de différents domaines d'intérêt qui ont accepté de nous aider pour cette tache de collecte de données. Nous avons en premier lieu sollicité ces utilisateurs à i) fournir, via un questionnaire, un ensemble de requêtes de recherche qui traduisent leurs centres d'intérêt, à ii) attribuer à chacune de ces requêtes des domaines d'intérêt à partir d'une liste prédéfinie de concepts qui sont extraits des niveaux supérieurs de l'ontologie Dmoz et à iii) regrouper ensemble les requêtes qui traduisent un même besoin d'information en leur attribuant un sujet de recherche parmi une liste prédéfinie de concepts qui sont extraits des niveaux inférieurs de cette ontologie. Le choix de ces deux listes de concepts est basé sur les deux hypothèses suivantes : d'une manière générale, une ontologie s'appuie sur une structure de données allant du plus général au plus spécifique. Ainsi les niveaux supérieurs de l'ontologie Dmoz peuvent représenter des domaines d'intérêt, et en allant vers le bas, les concepts sont plus spécifiques et peuvent donc représenter des sujets d'intérêt. Cela permet d'avoir une collection de requêtes qui soit pré annotée manuellement par de vrais utilisateurs. Elle aide à évaluer l'efficacité du modèle de classification de données exploité dans l'apprentissage du profil utilisateur. Deuxièmement, le choix de cette ontologie se justifie par la simple raison que les niveaux de représentation supérieurs du profil utilisateur sont construits à base de cette ontologie (cf. figure 4.1).

Les utilisateurs ont été amenés à remplir ce questionnaire 3 fois dans des périodes de temps différentes et séparées. Cela nous a permis d'avoir des besoins en information aussi variés que possible et évolutifs.

Nous avons recueilli 523 requêtes venant de 51 utilisateurs, chacune est annotée i) de plusieurs domaines qui représentent des concepts génériques de l'ontologie Dmoz, et ii) d'un sujet de recherche qui

représente un besoin d'information représenté par un concept spécifique de Dmoz. Cette étape est importante, elle nous a permis d'un côté de i) préparer au préalable les clusters de requêtes qui sont exploitées dans l'espace sémantique pour offrir des recommandations à base de sujets de recherche, puisqu'au début nous ne disposons toujours pas de données d'activités utilisateurs qui aident à offrir ces recommandations. Elle permet d'un autre côté de ii) connaitre à l'avance les sujets d'intérêt des utilisateurs qui nous ont été utiles pour extraire et préparer la collection de documents à indexer qui peut répondre aux différents besoins des utilisateurs. La qualité de la collection de documents offerte par le système aux utilisateurs est très importante pour une meilleure extraction de leurs données d'intérêt qui contribuent aux constructions de leurs profils.

Après chaque questionnaire, la liste des requêtes des différents utilisateurs est collectée et préparée puis les utilisateurs sont invités à utiliser le système avec les mêmes requêtes de recherche qui ont été fournies dans le questionnaire. Nous analysons ensuite les interactions de ces utilisateurs avec le système à travers l'interface de recherche/navigation et extrayons pour chacun les composantes qui constituent son profil, ces composantes s'énumèrent comme suit :

- **Une collection de requêtes.** Elle englobe la liste de i) toutes les requêtes de recherche qui sont soumises par l'utilisateur ainsi que la liste ii) des requêtes système qui sont prédéfinies sur l'interface (notamment au sein des deux facettes sémantique et identitaire) et ayant été exploitées par l'utilisateur pour affiner ou élargir ses recherches.
- **Une collection d'étiquettes**. Elle englobe la liste de i) toutes les étiquettes qui sont employées par l'utilisateur et ii) celles qui sont prédéfinies sur l'interface (notamment au sein de la facette sociale) et ayant été exploitées par l'utilisateur pour explorer/affiner les résultats de recherche.

Dans la représentation de ces deux collections de données, la distinction entre les requêtes/étiquettes qui sont employées par les utilisateurs et celles qui sont prédéfinies par le système et ayant été exploitées à travers l'interface de navigation par ces utilisateurs, est importante. Elle permet d'évaluer l'importance et l'efficacité des requêtes système au sein de l'interface proposée.

- **Les jugements de pertinence requête-document, ils sont notés par $Qrels_u$** : pour chaque requête un ensemble de documents pertinents est associé. Un document est considéré comme pertinent pour une requête de recherche s'il a été explicitement aimé par l'utilisateur ou a été annoté avec au moins une étiquette durant l'activité de recherche cible.

- **Les jugements de pertinence étiquette-document $\boldsymbol{Trels}_{\boldsymbol{u}}$**: pour chaque étiquette un ensemble de documents est associé. Un document est considéré pertinent pour une étiquette si i) cette dernière a été employée par l'utilisateur pour annoter le contenu de ce document ou ii) a été exploitée à travers l'interface de navigation pour localiser ce document (on parle des étiquettes qui sont prédéfinies par le système dans l'espace social pour l'exploration des résultats).
- **Annotation temporelle :** chaque donnée d'intérêt de l'utilisateur (requête, étiquette, document) est annotée temporellement avec une date de consommation qui aide à déduire sa fraicheur.
- **Annotation fréquentielle :** chaque donnée d'intérêt de l'utilisateur est annotée avec un score qui estime sa fréquence d'utilisation.
- **Facettes d'intérêt :** les données d'intérêt de l'utilisateur sont annotées chacune d'une facette qui représente la vue de données d'où la donnée a été consommée dans l'interface de navigation. Ceci aide à estimer les degrés d'intérêt de l'utilisateur pour les différentes facettes de données du système et évaluer l'utilisabilité de ces facettes au sein de l'interface de recherche. Cela permet également de concevoir une liste de jugements de pertinence pour chaque facette de données fe_i notée par $Qrels_{u,fe_i}$ qui aide à estimer la pertinence des résultats de cette facette. Cette liste associe à chaque requête q , émise par un utilisateur u, un ensemble de documents jugés pertinents par cet utilisateur au sein d'une facette de données fe_i.
- **Activités de recherche :** toutes les données d'intérêt de l'utilisateur (documents D_u, étiquettes T_u, requêtes système Q_u) qui sont exploitées derrière une requête de recherche q soumise à un instant « t » par l'utilisateur u, sont associées à une activité de recherche notée par A_u^t. Tel que : $A_u^t = \{q_u^t, Q_u^{A^t}, D_u^{A^t} T_u^{A^t}\}$.

Les caractéristiques de cette collection de tests sont illustrées dans le tableau 7.3.

Entité	#
Domaines d'intérêt généraux (extrais depuis 1er niveau Dmoz)	4
Domaines d'intérêt spécifiques (depuis niveau 2--3 et 4 Dmoz)	546
Sujets d'intérêt	78
Documents	112616
Requêtes /activités de recherche uniques	523
Longueur moyenne d'une requête	2,5
Documents pertinents dans les profils utilisateurs	31380
Étiquettes uniques dans les profils utilisateurs	20472
Utilisateurs	51
Facette d'intérêts uniques	3

Tableau 7. 3. La valeur correspondante pour chaque entité dans la collection de données étendue

VII.3.2. Stratégie d'évaluation

La stratégie d'évaluation consiste d'abord à instancier les modèles proposés avec des données réelles. Pour chacun de ces modèles, cala consiste à i) exploiter la collection de tests illustrée dans le tableau 7.3 pour mettre en œuvre un protocole d'implémentation puis ii) évaluer sa performance à travers des métriques d'évaluation. Chaque modèle exploite de cette collection de tests un sous-ensemble de composantes selon le besoin. Nous avons vu, dans la section 7.2, comment le modèle multi-facettes a été instancié au sein d'un prototype fonctionnel. Nous passons maintenant à l'évaluation de sa performance.

VII.4.Evaluation du modèle de la recherche d'information multidimensionnelle

Pour l'évaluation du modèle de la RI multidimensionnelle, deux critères d'efficacité sont considérés dans nos études, à savoir la pertinence des résultats de recherche et leur présentation sur l'interface utilisateur. Ces deux critères sont complémentaires pour l'efficacité d'un SRIF (cf. section 1.1.8).

VII.4.1. Efficacité des facettes de données et des valeurs de facettes

Nous nous intéressons à évaluer l'efficacité des facettes de données du système et celle des valeurs de facettes en termes de leur pertinence et leur utilisabilité dans l'interface de recherche. Cela est effectué en exploitant les profils utilisateurs qui sont construits à travers notre SRI. Pour cela, nous évaluons :

- Les degrés d'intérêt des utilisateurs pour ces facettes de données. Cela permet d'un côté d'évaluer l'utilité de ces facettes de données sur l'interface, et d'un autre d'enrichir les profils des utilisateurs

avec des valeurs de pertinence graduée envers ces facettes d'intérêt que nous jugeons utiles pour l'expérimentation du modèle personnalisé proposé dans le chapitre 5 qui réordonne les résultats selon les facettes d'intérêt des utilisateurs. Le degré d'intérêt d'un utilisateur u pour une facette de donnée fe_i a été défini dans l'équation 5.16. Il consiste à évaluer la proportion des données d'intérêt de l'utilisateur qui sont associées à une facette de données par rapport au contenu global de son profil. Ce degré est donné en pourcentage. Pour chaque facette, la valeur moyenne des pourcentages obtenus chez tous les utilisateurs est calculée (cf. figure 7.4).

- Nous évaluons la pertinence des résultats de recherche au sein de chaque facette fe_i estimée par les différents utilisateurs en exploitant les jugements de pertinence $Qrels_{u,fe_i}$. Cette pertinence est évaluée pour chaque requête au sein de chaque facette, puis la moyenne des pertinences obtenues de ces facettes, est calculée pour évaluer la pertinence globale du système. Celle-ci est relative à une seule requête. Donc, la moyenne des pertinences obtenues des différentes requêtes est calculée. Ces pertinences sont évaluées en termes de précisions (P@10, P@20 et MAP) et de rappel moyen. La précision P@X permet d'évaluer le degré de satisfaction des utilisateurs envers les X premiers documents résultants (cf. équation 2.12). La moyenne des précisions (MAP) exprime la capacité du système à répondre pertinemment à toutes les requêtes reçues par les différents utilisateurs et cela en considérant uniquement les points de précisions où un document est pertinent. La mesure du rappel permet de comparer la capacité de notre système à sélectionner plus de documents pertinents que les autres systèmes de référence. De la même manière, le rappel du système est calculé pour chacune des requêtes à différents points R10 et R20 (cf. page 43). Puis, la moyenne des valeurs de rappel, qui sont obtenues avec les différentes requêtes, est calculée. Les valeurs obtenues sont comparées à i) un SRI sémantique qui se base sur une indexation monodimensionnelle sans faire la distinction entre les types d'enrichissement qui sont appliqués lors de la description des documents et qui renvoie les résultats sur une seule facette de données, il comparé aussi à ii) un SRI traditionnel qui ne se base sur aucun type d'enrichissement où l'indexation et la recherche se basent uniquement sur le contenu identitaire de la requête et celui des documents dans l'index (cf. figure 7.7). Ces systèmes représentent les références de nos expérimentations, ils sont simulés en indexant les documents sur un seul espace d'indexation et en renvoyant les résultats sur une seule facette de données.

- Pour évaluer l'utilité des valeurs de facettes du système sur l'interface de navigation, nous évaluons leur fréquence d'utilisation au cours des activités de recherche des utilisateurs. Cela est effectué en calculant la proportion des requêtes système qui ont été exploitées par les utilisateurs par rapport au nombre total de leurs requêtes de recherche. Pour chaque utilisateur, cette fréquence est estimée, puis la valeur moyenne des fréquences obtenues chez les utilisateurs est calculée. Cette valeur est donnée en pourcentage (cf. figure 7.5).
- Pour évaluer la pertinence de ces valeurs de facettes, nous estimons la proposition des requêtes système qui sont exploitées par les utilisateurs et ayant pu offrir des documents pertinents pour ces utilisateurs, par rapport au nombre total des requêtes système dans leurs profils (cf. figure 7.6). Les requêtes système pertinentes sont issues des listes des jugements de pertinence des utilisateurs $Qrels_u$. Cela est basé sur l'hypothèse suivante : une valeur de facette est jugée pertinente pour une recherche donnée, lorsque son exploitation par l'utilisateur permet d'offrir des résultats pertinents.

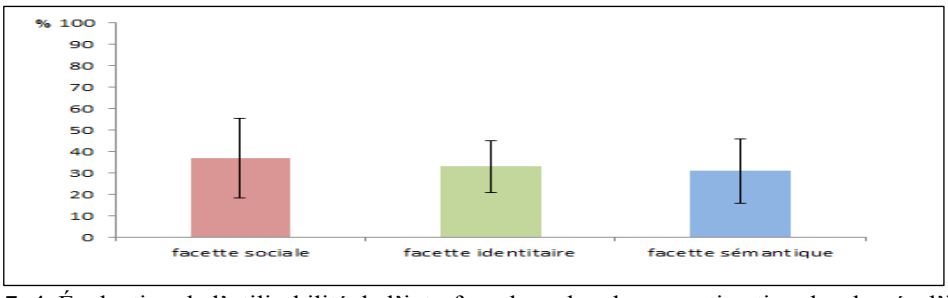


Figure 7. 4. Évaluation de l'utilisabilité de l'interface de recherche par estimation des degrés d'intérêt des utilisateurs envers les facettes de données

Comme nous pouvons le voir à travers la figure 7.4, les résultats montrent que, en moyenne, les utilisateurs sont intéressés de manière presque équitable par les trois facettes proposées. Plus concrètement, les trois facettes de données (social, identitaire et sémantique) ont obtenu respectivement des degrés d'intérêt de 37%, 33% et 31% par les utilisateurs du système.

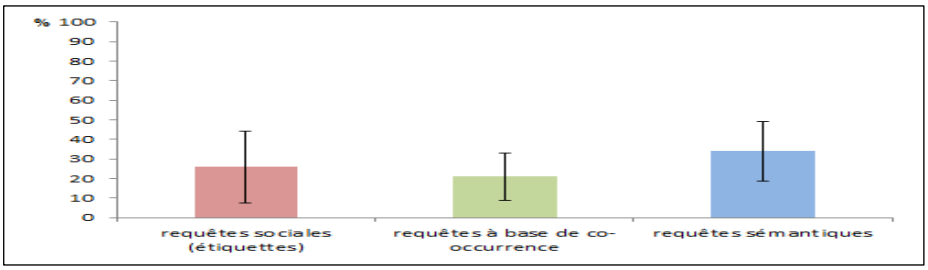


Figure 7. 5. Évaluation de l'utilisabilité de l'interface de recherche par estimation des fréquences d'utilisation des requêtes système (valeurs de facettes)

La figure 7.5 montre aussi l'intérêt des utilisateurs pour les différentes valeurs de facettes proposées. Plus concrètement, les requêtes sociales, sémantiques et celles basées sur le concept de cooccurrence ont été exploitées respectivement avec des fréquences d'utilisation de 26%, 34%, et 21% par les utilisateurs du système. Cela a permis d'alléger la charge cognitive de ces utilisateurs en leur proposons des requêtes prédéfinies sur l'interface de navigation pouvant répondre à leurs recherches. Pour évaluer la pertinence de ces valeurs de facettes, nous nous sommes intéressés à estimer uniquement la proportion des requêtes système ayant abouti à des résultats pertinents. Cette pertinence est centrée-utilisateur. C'est-à-dire, une requête système est estimée pertinente pour un utilisateur donné si ce dernier a jugé au moins un document pertinent lorsque cette requête est exploitée durant son activité de recherche. La figure 7.6 montre que parmi les requêtes système qui ont été exploitées par les utilisateurs respectivement 31%, 36% et 47% des requêtes identitaires, sociales et sémantiques ont permis de localiser des résultats pertinents. Nous pouvons donc conclure de ces résultats que nos perspectives fondées sur le concept de v-facettes (facettes basées sur les vues de données) sont optimistes.

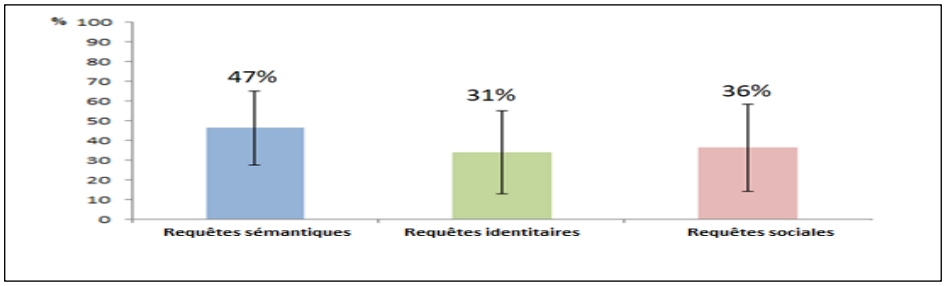


Figure 7. 6. Évaluation de la pertinence des valeurs de facettes par calcul de proportions de requêtes système dans les listes des jugements de pertinence $\boldsymbol{Qrels_u}$

Nous passons maintenant à l'évaluation de l'efficacité des facettes par pertinence de résultats de recherche. Les résultats illustrés par la figure 7.7 montrent que notre système est significativement meilleur qu'un système traditionnel qui ne se base sur aucun type d'enrichissement. Il présente également une amélioration considérable par rapport à un système qui se base pour l'indexation des documents sur un seul espace d'indexation et ne faisant aucune distinction entre les différentes interprétations possibles d'une requête.

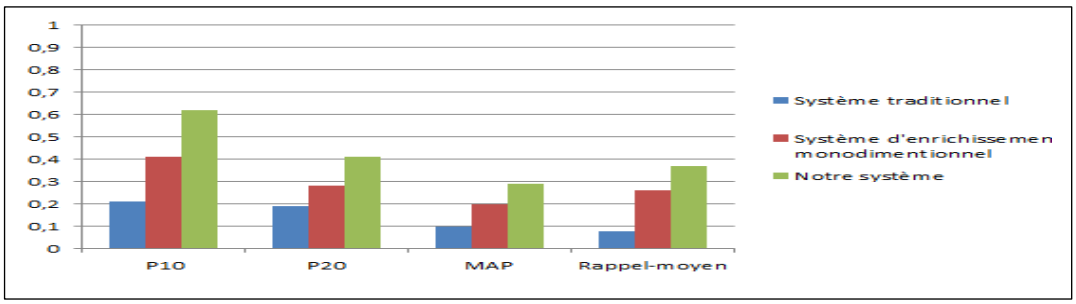


Figure 7. 7. Valeurs de précisions P10. P20, MAP et de rappel moyen dans un SRI traditionnel, SRI avec enrichissement monodimensionnel et un SRI multidimensionnel

Plus concrètement, les résultats montrent que par rapport au système traditionnel, notre système a obtenu des améliorations significatives de 195%, 115%, 190% et de 195% respectivement pour les précisions P10, P20, MAP et pour le rappel moyen, et de 51%, 46%, 45% et de 42% par rapport au système sémantique monodimensionnel. Cette amélioration est justifiée de la manière suivante :

- D'une part, une description qui se base sur plusieurs espaces d'interprétations permet au système d'appliquer différentes stratégies de recherche et d'affichage qui offre à l'utilisateur la possibilité d'accéder directement à l'espace souhaité, ce qui augmente la probabilité de sélectionner un résultat provenant de l'espace désiré et augmente la possibilité de trouver des résultats pertinents par rapport à une recherche monodimensionnelle.
- Un enrichissement sémantique permet d'augmenter la possibilité de localiser des documents dont le contenu identitaire ne couvre pas obligatoirement le contenu exact de la requête et cela en se basant sur les synonymes, hyponymes et hyperonymes du contenu de la requête ainsi que sur d'autres reformulations du besoin informationnel de l'utilisateur (le cas de l'enrichissement à base de sujet de recherche). Ce qui justifie une meilleure valeur du rappel système, c'est-à-dire, la capacité à trouver le plus de documents pertinents par rapport à un système traditionnel.
- Un enrichissement social permet d'impliquer les utilisateurs dans la description des documents. Ceci permet d'utiliser dans cette description le même langage employé par ces utilisateurs lors de la formulation de leurs requêtes de recherche. Cet enrichissement social augmente alors la possibilité de localiser des documents pertinents dans l'index documentaire. Ceci permet d'avoir une meilleure pertinence de résultats par rapport à un système traditionnel qui se base uniquement sur le contenu identitaire des documents pour leur représentation.

- D'autre part, même si un système se base sur un ou plusieurs types d'enrichissement, lorsqu'un seul espace d'indexation est utilisé pour décrire les documents sans faire la distinction entre les types d'enrichissement appliqués, cela limite les techniques qui peuvent être développées pour structurer les documents sur l'interface de navigation. Ainsi, suite à une requête de recherche, les résultats retournés sont mélangés dans un seul espace d'affichage et l'utilisateur doit filtrer et parcourir les résultats selon ses attentes. Prenons l'exemple d'un document web d_1 qui offre des recettes de tartes aux cerises. En plus des jetons qui constituent le contenu original de ce document, l'enrichissement sémantique permet de le décrire avec le jeton: fruit, puisque « cerise » est un hyponyme de « fruit ». Lorsque l'utilisateur est à la recherche de « recettes de tartes aux fruits », un système qui ne fait pas la différence entre les types d'enrichissement qui sont appliquées pour décrire les documents, renvoie tous les résultats mélangés ensemble sans faire la différence entre les documents qui couvrent la requête suite à un enrichissement de leur contenu ou par correspondance exacte avec leur contenu original. Ainsi, le document (d_1) est considéré autant pertinent qu'un autre document ayant le jeton « fruit » dans son contenu identitaire. Tandis qu'avec notre système, le document est classé par rapport aux autres documents de son espace de description et il est renvoyé dans une facette sémantique où les résultats sont basés uniquement sur la correspondance sémantique. Donc, si les attentes de l'utilisateur correspondent aux documents qui couvrent exactement sa requête, il accède directement à la facette identitaire. S'il souhaite explorer d'autres résultats sociaux (annotés par d'autres utilisateurs avec les jetons de sa requête) ou sémantiques offrant des recherches générique ou spécifique, il s'oriente vers les autres facettes, respectivement la facette sociale et sémantique.

VII.4.2. Évaluation du modèle de désambiguïsation de la requête utilisateur

L'efficacité d'un processus de recherche dépend principalement de l'interprétation de la requête utilisateur. Il est donc important de valider l'efficacité du modèle proposé dans notre étude pour désambiguïser le contenu d'une requête, notamment lorsque celle-ci contient un jeton polysémique. Pour ce faire, nous comparons les résultats obtenus lorsque cette tache de désambiguïsation est prise en considération par le système à ceux obtenus dans le cas contraire. Le modèle théorique proposé pour cette désambiguïsation se base sur le calcul d'une similarité composée (cf. section III.4.1.2). Cette similarité se

base sur l'exploitation d'une mesure de similarité et d'un dictionnaire sémantique. Il est donc important de choisir la méthode appropriée pour cette similarité. Nous effectuons des expérimentations où les trois mesures de similarité les plus connues dans la littérature sont exploitées, à savoir la mesure de Wu et Palmer, la mesure de Rada, et l'analyse sémantique latente (LSA). L'évaluation est effectuée en comparant les résultats obtenus par le système en réponse à un ensemble de requêtes ambiguës et en exploitant à chaque fois une mesure de similarité différente dans l'étape de désambiguïsation de la requête. Nous avons utilisé comme collection de tests la collection publique Dmoz qui classifie les pages web par catégories. La stratégie d'évaluation se résume par les points suivants :

- Indexer la collection de documents Dmoz dans un index monodimensionnel. Cette indexation est basée sur le contenu identitaire des documents. Nous jugeons ce type d'indexation suffisant pour cette évaluation, car le but est de tester l'efficacité du système à identifier la thématique d'un jeton polysémique dans une requête de recherche afin d'écarter les documents dans l'index qui ne couvrent pas le thème identité. L'évaluation est donc effectuée en termes de précision du système.
- Génération d'un ensemble de requêtes de recherche simulées à base des catégories Dmoz. Cette étape se résume par les étapes suivantes :
    - La sélection des catégories dont le titre est ambigu. Tel que Java, Apple, virus, stock, camp, bridge, sun, architecture, win, etc. Un titre est considéré comme ambigu lorsque son interprétation (la liste des documents qui lui sont associés dans l'ontologie dmoz) est liée à plusieurs thématiques (catégories) dans cette ontologie.
    - À partir de chaque catégorie sélectionnée, un ensemble de k-requêtes est simulé comme suit:
        - Extraire les documents qui sont associés à cette catégorie puis les repartir en k-groupes.
        - Chaque groupe de documents est représenté par un super document dont un vecteur de termes pondérés est préparé à partir de son contenu (notamment le contenu des titres et des descriptions des documents dans Dmoz). Ainsi, chaque catégorie est représentée par un ensemble de k-vecteurs pondérés.
        - Des requêtes sont simulées de chaque catégorie en i) sélectionnant les termes les plus pondérés dans les vecteurs générés puis ii) combinant chacun de ces termes au titre de la catégorie. Par exemple, pour la catégorie « virus » les termes les plus

pondérés qui sont extraits des documents sont : laptop, spyware, malware, infection, etc., et les requêtes simulées sont : virus laptop, virus spyware, virus malware, virus infection. Dans notre cas, la requête « virus infection » est éliminée puisque le jeton « infection » représente aussi à son tour un contenu ambigu, il ne permet pas donc de lever l'ambiguïté sur le mot virus. L'idée de notre modèle est d'utiliser les jetons qui ne sont pas ambigus dans la requête pour désambiguïser un jeton polysémique. Cette requête n'est donc pas adaptée pour tester le modèle.

- Préparation du fichier Qrels qui associe pour chaque requête un ensemble de documents pertinents. Il s'agit de relier chaque requête simulée à son groupe de documents à parti duquel cette requête a été simulée. De cette façon, l'évaluation du modèle se base sur une pertinence thématique. Nous jugeons cette pertinence comme étant suffisance et adéquate pour une telle évaluation puisque le but est de tester si le système est capable d'écarter les documents qui ne couvrent pas le thème auquel est liée la requête.
- L'ensemble de requêtes est utilisé pour interroger le système. Nous évaluons les résultats obtenus en réponse à ces requêtes en exploitant à chaque fois une mesure de similarité différente dans le processus de désambiguïsation.

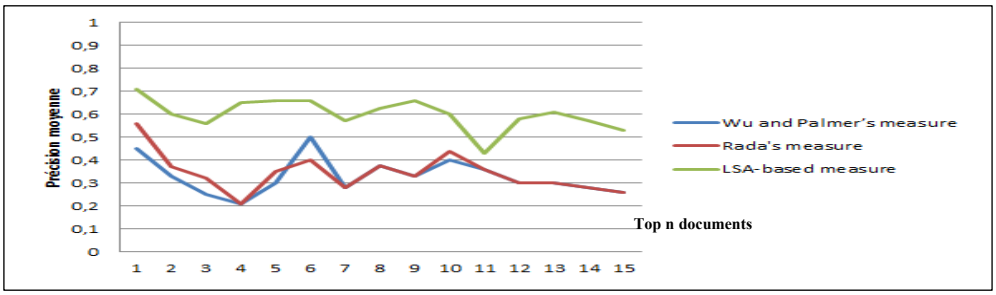


Figure 7. 8. Évaluation du modèle de désambiguïsation de la requête utilisateur selon trois mesures de similarité

Les résultats de la figure 7.8 montrent que la mesure de similarité LSA donne une meilleure précision des résultats par rapport aux autres mesures. Les résultats obtenus avec cette mesure sont comparés aux résultats qui sont obtenus dans le cas où la désambiguïsation de la requête n'est pas prise en compte.

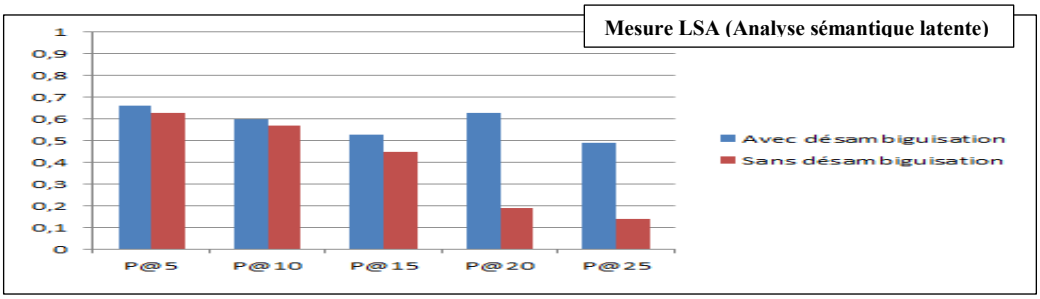


Figure 7. 9. Comparaison entre la RI avec et sans désambiguïsation du contenu de la requête utilisateur

La figure 7.9 montre une baisse de précision du système à partir des tops 20 documents lorsque les jetons polysémiques de la requête ne sont pas désambiguïsés. Cela peut être expliqué comme suit :

Les résultats d'une recherche sont les documents dont le contenu couvre un ou plusieurs jetons de la requête. Selon cette couverture qui peut aller d'une forte couverture (tous les jetons de la requête) à une faible couverture (un seul jeton de la requête), les documents sont classés du plus pertinent aux moins pertinents. Ceci explique la baisse de précision chez un système qui n'élimine les documents dont l'interprétation ne couvre pas la thématique de la requête et dont le contenu correspond à seul jeton de la requête en particulier celui qui peut être lié à plusieurs interprétations (le jeton polysémique). Par exemple, pour la requête « virus spyware», des documents qui traitent le virus de grippe peuvent être localisés puisqu'ils sont aussi décrits avec le jeton « virus ». Dans notre cas, la requête est désambiguïsée et son contenu est enrichi avec un concept qui représente le thème le plus approprié au jeton « virus » dans la requête. La requête exploitée par le système pour l'interrogation de l'index documentaire est la suivante : « virus spyware ». Elle est tokénisée en « virus », «spyware» puis enrichi en « virus informatique », « spyware ». Sachant que les documents dans l'index sont également décrits de cette manière, ainsi un document traitant le virus de la grippe est indexé avec le jeton « virus humain » au lieu de simplement « virus » (cf. page 109), ceci ne permet pas de le localiser lorsque la requête cible les virus informatiques.

VII.4.3. Mise à l'échelle

Nous présentons dans la figure 7.10, les résultats relatifs à des expériences de performance qui sont menées dans le but d'analyser le comportement du système lorsque: i) le nombre d'utilisateurs qui exploitent simultanément le système passe de 1 à 200 et i) le nombre de documents dans l'index

augmente jusqu'à 1 million 500 documents. Ces expériences ont été réalisées sur un PC avec 8 Go de RAM et un processeur Intel (R) Core (TM) i7-2670QM CPU@2.20GHz. Pour ce faire, nous avons eu recours au logiciel libre nommé JMeter qui permet d'effectuer des tests de performance en simulant le comportement de plusieurs utilisateurs sur une application web.

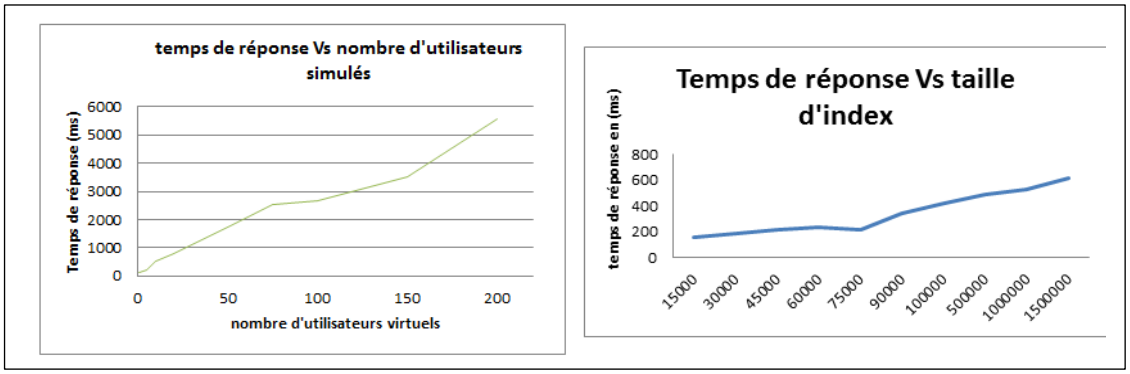


Figure 7. 10. Impact de l'augmentation de la taille d'index documentaire et du nombre d'utilisateurs sur le temps de réponse du système de recherche d'information

Les résultats montrent que les courbes évoluent presque linéairement, ce qui indique que la complexité du prototype est assez linéaire (O(N)). Par rapport aux spécifications de la machine utilisée dans ces expériences, les résultats indiquent que le système peut bien s'adapter à cette mise à l'échelle.

La section suivante présente la mise en œuvre et l'évaluation d'une nouvelle facette de données. Elle est proposée pour améliorer davantage la pertinence du système en personnalisant les résultats de recherche de chaque utilisateur selon ses centres d'intérêt. Ces données d'intérêt sont recueillies depuis les activités de recherche utilisateur et sont stockées dans son profil. Nous présentons comment ce profil est mise en œuvre et validé, puis évaluons par la suite la pertinence du système lorsque ce profil est intégré dans le processus de recherche d'information (RI).

VII.5. Évaluation du modèle de RI personnalisée

Dans notre étude, le modèle de la recherche d'information personnalisée (RIP) permet d'enrichir l'interface utilisateur avec une nouvelle facette qui offre des résultats personnalisés selon chaque utilisateur. Puisque ce processus consiste en premier lieu à la proposition d'un profil utilisateur qui modélise ses centres d'intérêt puis l'exploitation de ce profil dans la recherche d'information, son évaluation consiste donc à considérer deux étapes d'évaluation :

- L'évaluation de la qualité du profil utilisateur.
- L'évaluation de l'efficacité du modèle de RIP intégrant le contenu de ce profil.

VII.5.1. Évaluation de la qualité du profil utilisateur

Cette évaluation n'est pas nécessaire lorsque le profil de l'utilisateur est constitué uniquement de données qui sont saisies implicitement par cet utilisateur. Dans notre cas, l'évaluation de la qualité de ce profil est requise dans le but de tester la précision des données qui constituent les niveaux de représentation supérieurs de ce profil. Ces données sont automatiquement construites à partir des données d'activités spécifiques de l'utilisateur. Ces données spécifiques sont extraites depuis les interactions de l'utilisateur avec le système et constituent le niveau de représentation granulaire de son profil. Plus précisément, nous évaluons i) la précision des données conceptuelles du profil utilisateur constituées d'un ensemble d'activités de recherche où chacune est représentée par un ensemble de domaines d'intérêt, ii) la précision des données sémantiques de ce profil qui sont constituées d'un ensemble de sujets de recherche combinant les activités similaires en un seul centre d'intérêt, et iii) la précision des données au niveau le plus supérieur de ce profil qui englobe les groupes des sujets fortement connexes. Ces évaluations ont pour objectif de :

- Valider l'efficacité du processus de conceptualisation des données d'intérêt spécifiques aux utilisateurs en estimant la précision des domaines d'intérêt qui sont construits automatiquement à partir de ces données d'interactions pour représenter le profil de chaque activité de recherche.
- Valider l'efficacité du processus de délimitation de données d'activités par sujets d'intérêt en estimant la précision des sujets d'intérêt qui sont construits automatiquement par corrélation des activités de recherche successives de la couche sous-jacente.
- Valider l'efficacité du processus de construction des groupes de sujets fortement connexes (SFC) qui sont construits à base de détection de composantes connexes au sein d'un graphe.

Pour cette évaluation, plusieurs composantes sont nécessaires :

- Une collection de tests qui comprend :
    - Une collection de requêtes qui sont pré-annotées de sujets de recherche et de domaines de recherche.
- Dimension utilisateur qui représente les profils des utilisateurs. Chaque profil englobe :

- Les données d'activités de l'utilisateur (historique de navigations : requêtes, documents et étiquettes). Ces données représentent le premier niveau granulaire du profil utilisateur sur lequel se base la construction de ses niveaux de représentation supérieurs (cf. figure 7.12).
- Des jugements de pertinence $Qrels_u$ qui associent pour chaque requête de chaque utilisateur les documents pertinents.

➢ Un protocole d'évaluation qui se base sur la validation croisée. Cette technique de validation permet de créer les niveaux supérieurs du profil sur la base d'un ensemble d'activités d'apprentissage. Il consiste à répartir l'ensemble des activités de recherche de chaque utilisateur en un sous-ensemble d'apprentissage de k-1 activités, et la k^{éme} restante est gardée pour les tests du système avec le profil appris (cf. figure 7.11).

➢ Des métriques d'évaluation orientées sur le calcul de précision et de rappel du système.

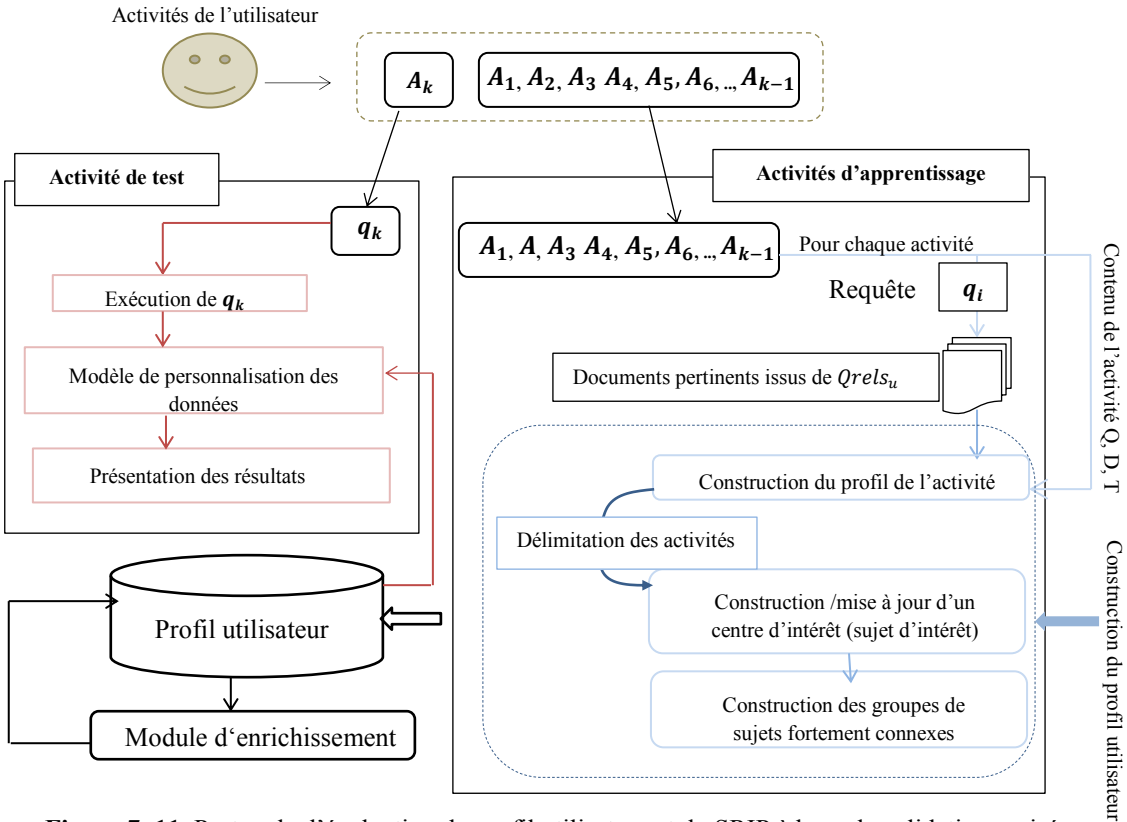


Figure 7. 11. Protocole d'évaluation du profil utilisateur et du SRIP à base de validation croisée

VII.5.1.1. Construction du profil utilisateur

La construction du profil utilisateur est effectuée en parcourant successivement la liste des activités d'apprentissage. A chaque itération, le profil de l'activité est construit et exploité pour la mise à jour du profil utilisateur par construction d'un nouveau centre d'intérêt ou par mise à jour d'un centre d'intérêt existant. Cette mise à jour du profil utilisateur est effectuée selon les deux cas discutés dans le modèle théorique et dépend principalement du processus de délimitation des activités de recherche qui compare leurs profils respectifs (cf. section IV.5.4).

La construction du profil d'une activité de recherche implique: i) la détection des domaines qui décrivent les documents pertinents de cette activité dans l'ontologie Dmoz (cf. section IV.5.4.1), les domaines résultants sont pondérés chacun d'un score qui représente son degré de correspondance sémantique avec le contenu de ces documents. ii) Ces domaines sont ensuite reliés les uns aux autres au sein d'un graphe conceptuel en se basant sur les liens de références dans l'ontologie Dmoz, puis iii) les données pertinentes de cette activité sont reliées au domaine le plus dominant dans le graphe pour définir un graphe topique de l'activité. Les profils des différentes activités de recherche constituent le niveau conceptuel du profil utilisateur.

Un centre d'intérêt est construit par combinaison de plusieurs profils d'activités qui sont liées à un même sujet de recherche. La liste de tous les centres d'intérêt (sujets d'intérêt) obtenus constitue à son tour le niveau sémantique du profil utilisateur qui vise à désambiguïser le contenu des données granulaires du profil utilisateur. En se basant sur une mesure de similarité entre graphes, les graphes conceptuels obtenus de chaque sujet sont reliés sémantiquement les uns aux autres au sein d'un graphe global G puis les sujets fortement connexes sont construits par construction de composantes fortement connexes au sein du graphe G à l'aide de l'algorithme de Kosaraju. Ces composantes constituent le niveau supérieur du profil utilisateur et représentent les clusters contextuels qui visent à construire des communautés d'intérêt. La figure 7.12 illustre les différentes étapes de cette construction du profil utilisateur.

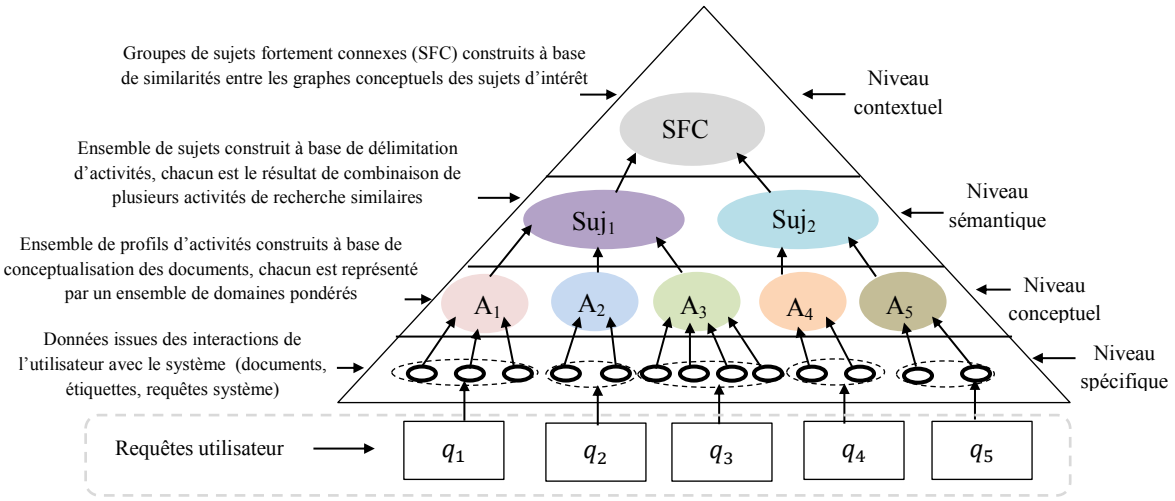


Figure 7. 12. Étapes de construction du profil utilisateur multi-niveaux

VII.5.1.2. Évaluation de la qualité des données conceptuelles du profil utilisateur

Le niveau conceptuel du profil utilisateur est constitué de l'ensemble des profils d'activités qui sont construits automatiquement depuis les données d'interactions de l'utilisateur, en projetant le contenu des documents pertinents issus du fichier $Qrels_u$ sur le contenu de Dmoz. L'évaluation de ce niveau de représentation consiste à estimer la précision des domaines qui sont associés à chaque profil d'activité lié à une requête de recherche, et cela en les comparants aux domaines de la collection de tests.

La précision à X domaines est alors calculée pour chaque activité de recherche en ordonnant les concepts par leur pondération du plus fort au moins fort où la valeur de X est variée de 1 à 10 (Daoud *et al.* 2010b). Puis la moyenne des précisions obtenues avec toutes les activités de recherche de tous les profils des utilisateurs est calculée à chaque point de précision.

$$P@X = \frac{nombre\ \ de\ domaines\ pertinents\ dans\ les\ X\ premiers\ rangs}{X}$$

Un domaine est considéré comme pertinent pour le profil d'une activité de recherche liée à une requête q_i, s'il correspond à une des annotations de cette requête dans la collection de tests.

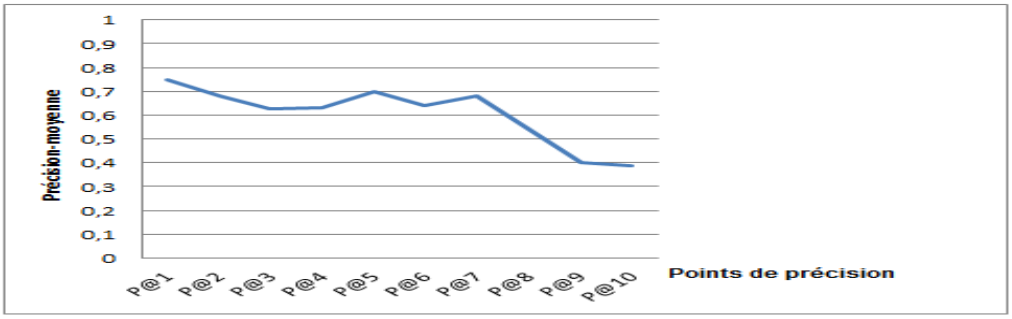


Figure 7. 13. Évaluation de la précision des données au sein des activités de recherche à X domaines d'intérêt

Les résultats illustrés dans la figure 7.13 montrent que le modèle d'apprentissage obtient 54% de précision moyenne sur les tops 8 domaines extraits des activités utilisateurs. Cette précision baisse au-dessous de 40% au 9eme domaine jusqu'à l'atteinte de 0.30 % au top 10 domaines. Ainsi, le nombre optimal des domaines considéré pour représenter une activité de recherche est défini à 8 domaines les mieux pondérés parmi la liste obtenue. Ce choix est justifié par le fait que le rang 9 est le point qui a marqué une chute au-dessous de 50% dans la précision des domaines au sein des activités d'apprentissage (P@9 =0.4).

VII.5.1.3. Évaluation de la qualité des données sémantiques du profil utilisateur

L'évaluation du niveau sémantique du profil utilisateur consiste à estimer la précision des sujets de recherche qui sont construits automatiquement par le système. Cette précision est liée au processus de délimitation des activités de recherche qui se base sur la corrélation de leurs profils respectifs construits au niveau conceptuel sous-jacent. Cette corrélation est fondée sur i) l'utilisation de la mesure de corrélation des rangs de Kendall et ii) la définition d'un seuil de corrélation optimal pour la détection de changement de sujets. Le choix de cette métrique est basé sur des expérimentations effectuées par d'autres travaux pour scruter le changement de contexte dans les taches de recherche des utilisateurs (Zemirli 2008) (Daoud *et al.* 2010b). Dans ces travaux, la métrique de Kendall a été utilisée pour comparer les rangs d'importance d'un ensemble d'éléments (concepts pondérés, termes pondérés, etc.) dans deux activités de recherche dans le but détecter une nouvelle tâche de recherche. Cette métrique a prouvé son efficacité par rapport aux autres mesures qui se basent sur une simple couverture entre ces

éléments, notamment la métrique webJaccard (Haveliwala *et al.* 2002) et la mesure du Cosinus. La métrique de Kendall permet de donner des valeurs de corrélation comprises dans l'intervalle de [-1, 1] où une valeur proche de −1 signifie que les activités sont complètement dissimilaires et une valeur proche de 1 signifie que les activités sont fortement similaires. Pour définir le seuil de corrélation, nous exploitons la collection des requêtes préannotées de sujets de recherche. Chaque ensemble de requêtes appartenant à un même sujet constitue une classe de requêtes (cf. figure 7.14).

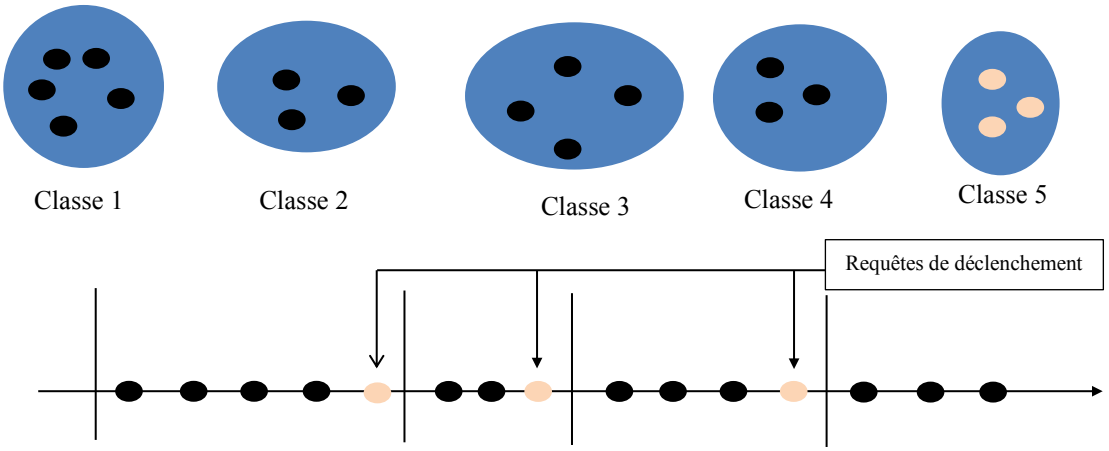


Figure 7. 14. Exemple de scénario d'apprentissage pour la délimitation des sujets d'intérêt des utilisateurs

Les requêtes d'apprentissage sont ordonnées chronologiquement pour simuler le comportement d'un utilisateur. Les requêtes qui marquent un changement de sujets sont connues à l'avance, nous les appelons par les requêtes de déclenchement. Une étape d'apprentissage se déroule en itérons sur les requêtes d'apprentissage et évaluons à chaque itération « i » la corrélation entre le profil de l'activité courante et le centre d'intérêt construit à l'itération précédente « i-1».

Cette étape est répétée en faisant varier la valeur du seuil de -1 à 1 et estimons pour chaque valeur l'efficacité du modèle d'apprentissage à classer les requêtes dans leurs classes adéquates. Cette efficacité est estimée en termes de précision et évaluée selon les deux métriques P1 et P2 suivantes (Daoud *et al.* 2010b).

$$P1_{classification\ par\ sujets} = \frac{nombre\ de\ requetes\ correctement\ classées\ au\ sein\ des\ sujets\ \ prédéfinis}{nombre\ \ de\ requetes\ au\ sein\ de\ la\ collection\ d'apprentissage}$$

$$P2_{déclanchement\ nouveau\ sujet} = \frac{nombre\ de\ requetes\ de\ déclanchement\ correctement\ classées}{nombre\ \ de\ requetes\ de\ déclanchement\ au\ sein\ de\ la\ collection\ d'apprentissage}$$

Les métriques P1 et P2 permettent respectivement d'évaluer l'efficacité du processus d'apprentissage à détecter les requêtes qui sont liées à un même sujet, et à détecter les changements des sujets.

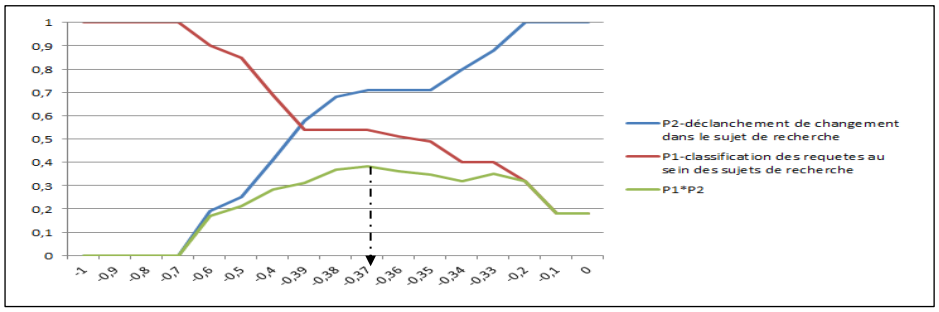


Figure 7. 15. Définition de la valeur optimale pour la délimitation des sujets d'intérêt des utilisateurs à base de calcul des produits (P1*P2) des précisions obtenues avec les deux critères de pertinence P1 et P2

La valeur optimale du seuil de délimitation représente le meilleur compromis entre les deux critères d'efficacité. Il s'agit de la valeur qui donne un résultat élevé pour les deux précisions P1 et P2. Pour définir cette valeur, nous adoptons deux solutions comme suit :

1. Pour chacun des deux critères P1 et P2, nous classons dans un ordre croissant les valeurs des précisions obtenues puis attribuons à chacun un rang $rang_i$ qui dénote son classement par rapport aux autres valeurs. Pour chaque valeur de seuil, la somme des deux rangs, qui sont obtenus des deux critères, est calculée. Le seuil ayant la somme maximum est celui qui présente le meilleur compromis entre les deux critères. La figure 7.16 indique que le seuil optimal est de -0.37. Afin de valider cette valeur, nous adoptons une deuxième solution.
2. Puisque l'image des deux métriques est comprise entre 0 à 1 (positive), nous nous basons sur la multiplication des précisions. Ainsi $ArgMax_{\Omega}(P1)$ ET $ArgMax_{\Omega}(P2)$ = argMax (P1*P2). La figure 7.15 montre que la valeur qui maximise la multiplication est égale à - 0.37. Cette valeur est donc retenue pour marquer le changement de sujets entre les activités des utilisateurs.

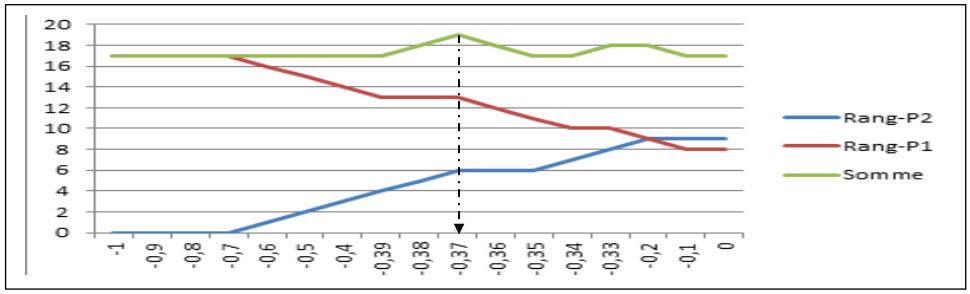


Figure 7. 16. Définition du seuil de corrélation optimal à base de calcul des sommes de rangs d'importances des précisions obtenues avec les deux critères de pertinence P1 et P2

VII.5.1.4. Évaluation de la qualité des données contextuelles du profil utilisateur

Ce niveau de représentation englobe les groupes de sujets fortement connexes (SFC) qui sont construits automatiquement depuis la couche sémantique des profils des utilisateurs. Ce niveau a pour objectif de construire des communautés d'intérêts qui partagent des sujets connexes. Cet aspect de communautés d'intérêts est exploité dans notre étude au sein de divers processus que nous énumérons ci-après.

À cause de l'indisponibilité d'une ressource de référence qui nous permet de valider la précision des groupes de sujets construits, nous évaluons l'efficacité de ce niveau de représentation lors de son intégration au sein du processus de recommandation de données (cf. section VII.5.2) et au sein du modèle de la recherche personnalisée (cf. section VII.5.3). Cela est basé sur l'hypothèse que la qualité du profil utilisateur a un impact direct sur la performance du modèle qui exploite son contenu. L'efficacité de ce niveau de représentation dépendra ainsi de :

- La capacité du modèle de recommandation à optimiser le processus d'extraction de corrélations entre les données d'intérêt des utilisateurs en délimitant ces données par groupes de sujets fortement connexes (cf. section VII.5.2.1).
- La capacité de ce modèle de recommandation à personnaliser les règles d'association pour les utilisateurs au cours de l'inférence de données en se basant sur i) la classification de ces règles en classes contextuelles où chacune est issue d'un groupe de SFC, ii) l'affectation d'un groupe SFC le plus représentatif des intérêts de chaque utilisateur, puis iii) la sélection personnalisée des règles intéressantes relatives au groupe d'intérêt sélectionné (cf. section VII.5.2.4).

- La capacité du SRIP à améliorer la pertinence des résultats lorsque ce système se base sur les groupes de sujets pour définir les utilisateurs voisins et intègre leurs intérêts pour améliorer leurs recherches (cf. section VII.5.3.1).
- La capacité du SRI à identifier le besoin informationnel de l'utilisateur lorsque sa requête est ambiguë et liée à plusieurs sujets de recherche dans son profil (cf. section VII.5.32).

La section VII.5.2 a permis de valider la précision du contenu multidimensionnel du profil utilisateur. La section qui suit présente l'évaluation d'un modèle exploitant ce profil pour enrichir davantage son contenu au sein d'un processus de recommandation par inférence collaborative de données.

VII.5.2. Évaluation du module d'enrichissement du profil utilisateur par recommandation collaborative d'intérêts

L'évaluation du module d'enrichissement du profil utilisateur consiste à évaluer les résultats obtenus par le processus de recommandation de données. Ce processus exploite les profils des utilisateurs pour leur inférer de nouveaux intérêts. Cette exploitation permet d'estimer l'efficacité des différents niveaux multidimensionnels de ces profils au sein de ce processus de recommandation. Cette efficacité est estimée au cours de différentes étapes de cette recommandation de données comme suit :

Étape 1. Lors de l'extraction des règles d'association qui alimentent ce processus de recommandation de données. Cela est effectué en évaluant la capacité du modèle proposé à alléger le temps nécessaire pour extraire les données d'activités fréquentes des utilisateurs lorsque :

i) ces données sont préparées au départ en groupes de sujets fortement connexes dans le but de réduire le nombre de données traitées,

ii) et lorsque ces données sont structurées à l'intérieur de chacun de ces groupes en sujets de recherche et cela dans le but d'éliminer d'abord les sujets non fréquents avant d'extraire les données d'activités granulaires fréquentes, ce qui peut aider à réduire encore plus le nombre de données traitées.

Étape 2. Cette efficacité est estimée aussi lors du processus d'inférence de données, en évaluant la capacité de ce processus à améliorer la pertinence de résultats de recommandation à travers les trois techniques proposées:

- Lorsque l'ambigüité est levée sur les données d'intérêt polysémiques des utilisateurs en se basant sur leurs hauts niveaux de représentation.
- Lorsque la corrélation entre les utilisateurs est extraite à travers les hauts niveaux de représentation de leurs intérêts, notamment lorsque les utilisateurs ne possèdent pas les mêmes données d'activités granulaires, mais partagent les mêmes sujets d'intérêt. Ceci aide à améliorer le rappel du système, c'est-à-dire, sa capacité à repérer plus de documents pertinents pour les utilisateurs.
- Lorsque la sélection des règles d'association est personnalisée pour les utilisateurs en se basant à la fois sur i) la classification de ces règles en classes contextuelles et ii) l'identification de la classe la plus pertinente pour chaque utilisateur, c'est-à-dire celle qui représente le plus ses intérêts courants.

VII.5.2.1. Évaluation du processus d'extraction des itemsets fréquents

Pour l'évaluation de ce processus, nous suivons le protocole d'évaluation suivant :

1. Il est important avant tout d'augmenter la base d'activités des utilisateurs, car la collection de données que nous avons extraite depuis notre système est petite et ne présente pas beaucoup de corrélations entre les utilisateurs, elle n'est donc pas adéquate pour de tels tests. La construction de cette nouvelle base de tests étendue est fondée sur la création de requêtes simulées et de créations de jugements de pertinences à partir de la collection de données de départ Delicious. Le processus d'extension de cette collection est expliqué dans l'annexe 5.
2. Construire les profils des utilisateurs depuis la nouvelle collection.
3. Nous passons ensuite à l'évaluation du processus en question qui commence par la validation de la pertinence des hypothèses sur lesquelles se base cette extraction de données fréquentes. Avant de présenter ces hypothèses, il est important de faire un rappel sur quelques notions de base que nous jugeons importantes pour la compréhension de la suite de nos évaluations.

   Définitions. Un itemset est un ensemble d'items. Un item peut représenter n'importe quel objet de contenu, dans notre cas il peut être une donnée granulaire (étiquette ou requête) ou une donnée générique (un sujet d'intérêt). Donc un itemset spécifique est un ensemble d'items granulaires (ensemble d'étiquettes ou de requêtes) et un itemset générique est un ensemble de sujets d'intérêt. Un k-itemset est un itemset (générique /spécifique) constitué de k-items (génériques /spécifiques).

Les hypothèses sur lesquelles se base ce processus d'extraction d'itemsets fréquents s'énumèrent comme suit :

- **Hypothèse 1-**les étiquettes d'annotation des utilisateurs sont parfois personnelles et non compréhensibles et ne permettent pas d'extraire beaucoup de corrélations entre les utilisateurs. L'intégration des requêtes de recherche dans la préparation de données granulaires des utilisateurs peut être utile pour augmenter la corrélation entre les utilisateurs.
- **Hypothèse 2-**les sujets d'intérêt non fréquents englobent les données d'activités granulaires non fréquentes. Cette hypothèse est exploitée pour éliminer d'abord les sujets non fréquents avant de traiter les données d'activité granulaires. Ceci permet de supprimer les calculs inutiles.
- **Hypothèse 3-**lorsqu'un k-itemset générique (tel que k > 1) est non fréquent, cela implique que tous les k-itemsets spécifiques obtenus par combinaison de leurs items spécifiques respectifs ne sont pas fréquents aussi. Ceci permet de réduire le temps de calcul en supprimant les calculs inutiles.
- **Hypothèse 4-**les itemsets fréquents appartiennent généralement à des sujets fortement connexes, ainsi les itemsets génériques constitués de k sujets non connexes (tel que k>1) ne sont pas fréquents. C'est la raison pour laquelle les données de départ sont délimitées en k-groupes de SFC. De cette façon, en se basant sur l'hypothèse 3, chaque groupe de sujets est traité à part, ceci aide à réduire le nombre d'itemsets traités de 2^N à $2^{N_1}+...+2^{N_k}$ pour les k-groupes de sujets, où N est le nombre de données d'activités granulaires de départ, tel que $N = N_1 + \ldots + N_k$.

4. Une fois les hypothèses sont validées, nous estimons le temps nécessaire avec notre modèle pour l'extraction des itemsets spécifiques fréquents où les sujets de recherche et les groupes de sujets connexes sont utilisés pour préparer les données de départ. Nous le comparons ensuite au temps de calcul nécessaire avec le modèle classique où les données de départ représentent les données d'activités granulaires des utilisateurs (les étiquettes et les requêtes de recherche).

A. Validation des hypothèses

Hypothèse 1. La validation de la première hypothèse consiste à analyse les corrélations entre les utilisateurs lorsque les requêtes et les étiquettes sont prises en compte ensemble pour préparer les données d'activités granulaires des utilisateurs. Ces données font l'objet de l'extraction d'itemsets spécifique

fréquents. Nous commençons par présenter les motivations de cette hybridation ainsi que la technique adoptée pour cette préparation de données, puis nous passons à la validation de l'hypothèse en question.

Motivations. Les requêtes de recherche des utilisateurs peuvent contribuer à l'augmentation de corrélations entre les activités des utilisateurs surtout que les étiquettes qui sont employées par ces derniers sont généralement personnelles et réduisent la possibilité de détecter des corrélations intéressantes pour la recommandation collaborative. Cela est basé sur l'hypothèse suivante : lorsque l'utilisateur est à la recherche de documents à travers un moteur de recherche, il fait plus d'effort dans la formulation de sa requête que lorsqu'il annote des documents pour les sauvegarder dans sa liste favorite. Ainsi, le langage de termes utilisé par un utilisateur lors de ses recherches sur le système est d'un côté moins ambigu et plus compréhensible que celui exploité lors de ses annotations, et d'un autre il peut présenter plus de corrélation avec le langage des autres utilisateurs, car les utilisateurs cherchent à cibler les mêmes documents au sein d'un même système. Ainsi, la considération de ces requêtes de recherche peut être utile pour les utilisateurs qui consomment les mêmes documents, mais les annotent différemment. Ceci permet de surmonter les faiblesses des étiquettes que nous avons soulevées dans la section IV.4 page 137.

En outre, la considération de ces requêtes dans cette extraction de corrélations peut être aussi utile lorsque les utilisateurs ne sont pas socialement interactifs, c'est-à-dire, ils se limitent uniquement à la soumission des requêtes de recherche et la consommation de ressources. L'historique d'annotations de ces utilisateurs est considéré comme faible et pas insuffisant pour détecter des corrélations avec les autres utilisateurs.

Technique d'hybridation. Comme expliqué précédemment dans le modèle d'interprétation de la requête utilisateur (cf. section VII.2.3), une requête de recherche est composée de plusieurs jetons. Ainsi, l'ensemble de jetons constituant le contenu des requêtes de recherche de chaque utilisateur est combiné avec l'ensemble de ses étiquettes pour représenter ses données d'activités de départ. Puisque le but est d'augmenter les corrélations entre les utilisateurs, ces deux ensembles ne sont pas considérés séparément, mais ils sont pris en compte en sein d'une seule transaction. Chaque utilisateur est donc représenté par un ensemble de jetons qui peuvent être des étiquettes ou des requêtes.

Validation. Pour valider cette hypothèse, nous comparons le nombre d'itemsets spécifiques fréquents qui peuvent être extraits lorsque les étiquettes ou les requêtes des utilisateurs sont prises en

considération à la fois seules durant cette étape d'extraction, et le cas où les deux catégories de données sont combinées ensemble. Pour cela chaque utilisateur est représenté par la transaction de ses activités, et un support minimum de 0.5 est exploité pour extraire les itemsets fréquents, cette valeur est définie expérimentalement (cf. section VII.5.2.2).

support min = 0.5	Prise en considération uniquement des étiquettes	Prise en considération uniquement des requêtes	Combinaison des étiquettes et des requêtes
Nombre d'itemsets spécifiques fréquents	522	187	961

Tableau 7. 4. Nombre des itemsets spécifiques fréquents extraits lorsque plusieurs techniques sont appliquées pour préparer les données d'activités des utilisateurs

Nous pouvons voir à travers le tableau 7.4 que le nombre de corrélations (les itemsets fréquents) extraites des intérêts des utilisateurs lorsque les requêtes sont combinées aux étiquettes présente une amélioration par rapport au cas où les deux ensembles sont pris en compte chacun seul. Cela permet de valider cette idée d'hybridation.

Hypothèses 2 et 3. Il est important de mettre l'accent sur le fait que l'objectif de cette étape d'évaluation n'est pas d'améliorer l'ensemble des itemsets spécifiques fréquents qui peuvent être extraits de notre modèle, mais de valider les hypothèses posées dont l'objectif est d'optimiser le temps nécessaire pour cette étape d'extraction par rapport au modèle classique. Ainsi, les résultats obtenus par le modèle classique sont exploités comme base de référence pour tester l'efficacité de ces hypothèses. Pour valider les deux hypothèses (2 et 3), nous nous basons sur une comparaison qui consiste à vérifier si les itemsets spécifiques qui sont éliminés par notre méthode sont considérés aussi comme non fréquents avec la méthode classique. Et si notre modèle arrive à éliminer tous les itemsets spécifiques qui sont considérés comme non fréquents par le modèle classique. Pour ce faire, nous estimons la précision de deux ensembles de données. Le premier est l'ensemble des itemsets spécifiques qui ont été correctement éliminés par notre modèle, le deuxième ensemble est celui des itemsets spécifiques qui ont été mal éliminés. Pour cela, nous nous basons sur la table de contingence illustrée dans le tableau 7.5 sur laquelle deux métriques sont définies.

Par le modèle classique

Par notre modèle	Non fréquents	Fréquents
Éliminés	Vrais positifs	Faux positifs
Non éliminés	Faux négatifs	Vrais négatifs

Tableau 7. 5. Table de contingence pour l'évaluation de la pertinence du processus d'extraction d'itemsets spécifiques fréquents de notre modèle

- **Vrais positifs** : Données non fréquentes correctement classées. Il s'agit des données qui sont considérées comme non fréquentes avec les deux modèles.
- **Faux positifs** : Données fréquentes mal classées. Ce sont les données qui sont considérées comme non fréquentes par notre modèle et elles sont donc éliminées, tandis qu'elles sont fréquentes par le modèle classique.
- **Vrais négatifs** : Données fréquentes correctement classées. Il s'agit des données qui sont considérées comme fréquentes avec les deux modèles.
- **Faux négatifs** : Données non fréquentes mal classées. Ce sont les données considérées comme fréquentes avec notre modèle, elles ne sont donc pas éliminées, tandis qu'elles sont non fréquentes avec le modèle classique.

$$Précision = \frac{vrai\ positifs}{vrai\ positif + faux\ positifs} = \frac{Nombre\ d'itemsets\ nonfréquents\ bien\ classées\ (élliminés)}{Nombre\ d'itemsets\ élliminés}$$

$$Rappel = \frac{vrai\ positifs}{vrai\ positif + faux\ négatifs} = \frac{Nombre\ d'itemsets\ nonfréquents\ bien\ classée\ (éliiminés)}{Nombre\ d'itemsets\ nonfréquents\ existants}$$

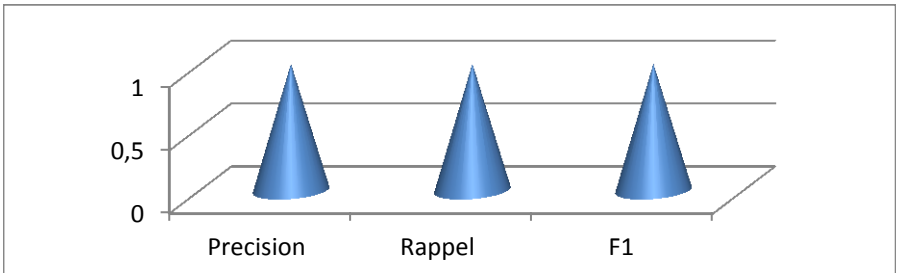


Figure 7. 17. Valeurs de précision, de rappel et de compromis F1 du nouveau modèle d'extraction des données d'activités fréquentes des utilisateurs par rapport au modèle classique

Les résultats de la figure 7.17 sont satisfaisants, ils montrent que le processus d'extraction des itemsets spécifiques fréquents qui se base sur plusieurs niveaux de représentation de données, en commençant du général (les sujets d'intérêt) vers le spécifique (les données d'activités granulaires) obtient les mêmes résultats qu'avec le modèle classique.

Hypothèse 4. La validation de la quatrième hypothèse consiste à analyser les corrélations entre les sujets d'intérêts dans les profils des utilisateurs. Plus précisément, il s'agit i) d'extraire les k-itemsets génériques fréquents lorsque les sujets ne sont pas délimités, puis ii) vérifier si aucun k-itemsets générique fréquent obtenu (tel que k>1) n'est constitué de k sujets non connexes. Cette analyse est effectuée en faisant varier le support minimum de 0.1 à 0.9. Nous analysons la proportion de ces itemsets fréquents par rapport à l'ensemble total des itemsets génériques fréquents.

Les expérimentations ont obtenu un pourcentage de 4% d'itemsets génériques fréquents constitués de k sujets appartenant à différents groupes. Ce pourcentage est obtenu avec un seuil de fréquence de 0.1. Lorsque ce seuil de fréquence est augmenté à 0.2, aucune corrélation entre les différents groupes n'est extraite. Nous estimons que le seuil de fréquence avec lequel la corrélation a été extraite est faible. En outre, le pourcentage de ces itemsets est très faible aussi, ce qui permet de valider notre hypothèse. Ainsi, les 4 hypothèses proposées sont validées et peuvent être utilisées durant le reste des expérimentations.

food, music, video
food, new, music
food, design, video
healthy food, laptop
programming, fitness sport
food, photography, architecture, art
history, programming ,art, music
google, tv, food

Tableau 7. 6. Exemple d'itemsets fréquents constitués d'items appartenant à différents groupes de sujets

La délimitation de données permet de créer des communautés d'intérêts à base de groupes de sujets connexes. Chaque communauté englobe les utilisateurs ayant un intérêt pour un ou plusieurs sujets du groupe. Un autre avantage qu'on peut soulever de cette délimitation est que le nombre de transactions traitées (le nombre d'utilisateurs) est réduit dans chaque communauté d'intérêt. Ainsi, lorsque la fréquence des itemsets est calculée à travers la métrique du support, cela peut aider à détecter certains itemsets fréquents qui ne peuvent pas être extraits lorsque le nombre total des interactions est grand, c'est-

à-dire, lorsque tous les utilisateurs sont pris en compte au sein d'une seule communauté globale. Cela est justifié par le fait que le support d'un itemset dépend principalement du nombre total de transactions dans la base de données (cf. définition II.1). Ainsi, lorsque ce nombre est réduit, la fréquence augmente.

En outre, nous jugeons que la fréquence d'un itemset est plus pertinente lorsqu'elle est estimée au sein d'une communauté d'intérêts avec une grande connectivité sémantique, que lorsqu'elle est estimée au sein de toutes les communautés qui peuvent englober différents intérêts (connectivités faibles). La pertinence de ses itemsets est évaluée dans la section VII.5.2.2 lors de l'extraction des règles d'association à travers les trois métriques de support, de confiance et de lift.

B. Évaluation de l'efficacité du modèle à réduire le temps de calcul des itemsets fréquents

Afin de tester l'efficacité du modèle proposé à réduire le temps de calcul, nous l'avons testé sur 3 groupes de données. Le premier groupe représente les données d'activités qui ne présentent pas de corrélations (ni spécifique ni générique) les unes aux autres, le deuxième groupe est l'ensemble d'activités qui présentent les unes aux autres des corrélations génériques, mais pas de corrélations spécifiques, le dernier groupe est l'ensemble des données d'activités qui englobe les données avec corrélations génériques et spécifiques. Cela permet d'évaluer notre modèle dans différents cas de corrélations. Pour ce faire, nous suivons le processus illustré dans la figure 7.19, puis exploitons les groupes de données que nous nommons dans ce processus par « groupe-1 de données », « groupe-4 de données » et « groupe 2 de données ». Comme nous pouvons le voir à travers ce processus, les groupes de données choisis correspondent aux critères de corrélations précitées. Nous présentons, à travers la figure 7.18, les résultats obtenus avec notre modèle en termes de gain en temps de calcul par rapport au modèle classique. Les résultats montrent que notre modèle présente une amélioration considérable dans tous les cas étudiés sauf dans le cas où seules des corrélations génériques existent entre les données des utilisateurs sans de corrélations spécifiques.

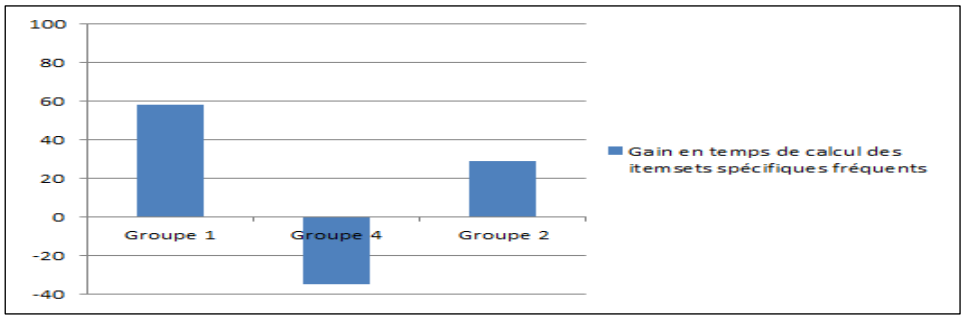


Figure 7. 18. Gain en temps de calcul des itemsets fréquents avec notre modèle par rapport au modèle classique au sein de différents cas de corrélations entre les intérêts des utilisateurs

Ces résultats peuvent être expliqués comme suit :

- Dans le cas où il n y a pas beaucoup de corrélations entre les activités granulaires des utilisateurs, notre modèle peut être un peu plus lent, car le processus traite d'abord les données génériques des utilisateurs (leurs sujets d'intérêt) puis les données granulaires (leurs historiques d'étiquettes et de requêtes). Plus particulièrement, dans certaines situations un sujet ou un ensemble de k sujets sont fréquents alors que les données granulaires ne le sont pas, c'est le cas où les utilisateurs ont des sujets d'intérêt commun, mais n'emploient pas les mêmes requêtes et étiquettes. Cette situation génère un temps de calcul supérieur à celui obtenu avec le modèle classique, car les calculs effectués par notre modèle sur les données génériques sont si on peut dire inutiles. De cela, on peut déduire qu'un sujet fréquent n'implique pas toujours la fréquence des données d'activités granulaires.
- Dans le cas où les corrélations spécifiques ne sont pas présentes entre les données, le traitement de ces données par leur haut niveau de représentation permet d'éliminer les sujets non fréquents, ce qui permet de supprimer les calculs inutiles effectués par le modèle classique sur l'ensemble de données granulaires qui sont liés aux sujets non fréquents éliminés. L'élimination d'un k-itemset générique peut éliminer plusieurs k-itemsets spécifiques. Ceci explique l'optimisation dans le temps de calcul.

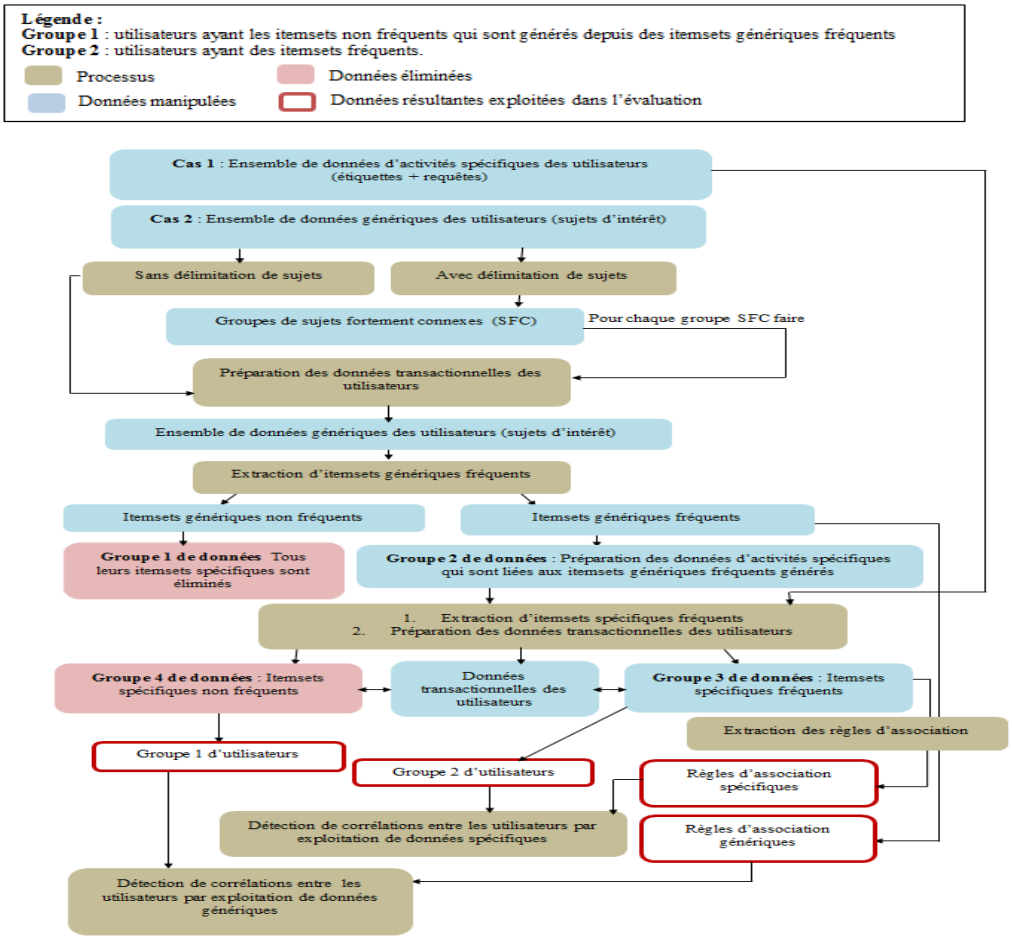


Figure 7. 19. Stratégie d'évaluation du système d'inférence d'intérêts des utilisateurs

Nous soutenons les résultats expérimentaux obtenus avec des démonstrations théoriques de deux cas d'étude (cf. Annexe 4). La première démonstration représente le cas où le nombre de calculs effectués par notre modèle pour extraire les itemsets spécifiques fréquents est amélioré de 32% par rapport au modèle classique. Dans le deuxième cas, les sujets des utilisateurs sont fréquents et les données d'activités spécifiques qui constituent ces sujets ne le sont pas. Ceci augmente le nombre d'opérations qui sont traitées par notre modèle. Cette augmentation peut être plus intense lorsque le nombre d'utilisateurs est grand et lorsqu'il y a beaucoup de sujets d'intérêt à traiter.

La section prochaine présente le processus qui exploite les itemsets fréquents obtenus de cette étape pour extraire les règles d'association du système.

VII.5.2.2. Extraction des règles d'association

L'extraction des règles d'association consiste à exploiter les itemsets fréquents obtenus pour les représenter sous forme d'inférences (A→B) que le système exploite pour recommander des données d'intérêts aux utilisateurs. Le but de cette étape d'évaluation est de trouver toutes les inférences qui satisfirent certaines restrictions de support et de confiance. Le support permet de mesure la fiabilité de la règle (cf. définition II.1). Plus la valeur de ce support est grande, plus les règles extraites sont moins nombreuses et évidentes, mais moins utiles pour l'utilisateur. Il est donc nécessaire de définir une valeur de support optimale qui soit suffisamment bas pour extraire des informations importantes. Ceci risque par contre de générer une quantité importante de règles, ce qui rend difficile leur analyse. La confiance permet à son tour d'évaluer la précision d'une règle, elle est exploitée pour éliminer parmi les associations fréquentes, celles les moins importantes (cf. définition II.2). Pour trouver les valeurs optimales de support minimum et de confiance minimale, nous nous sommes basés sur un troisième indicateur de pertinence des règles qui dépasse le support et la confiance qui est le Lift (cf. équation 2.18). Une règle est jugée pertinente lorsque son lift est supérieur à 1(Agrawal *et al.* 1993). Ainsi, nous faisons varier des valeurs de chacune des deux métriques de support et de confiance de 0 à 1, puis évaluons le Lift de chaque règle obtenue à un point combinaison de valeurs (support, confiance). La combinaison qui obtient plus de règles ayant un Lift supérieur à 1 est celle qui sera retenue. Le système a extrait 1047 règles d'association spécifique et 279 règles génériques ayant un Lift supérieur à 1, avec un support égal à 0,5 et une confiance égale à 0,6.

VII.5.2.3. Évaluation du processus d'inférence de données

Cette évaluation consiste à tester l'efficacité du système d'inférence de données à :

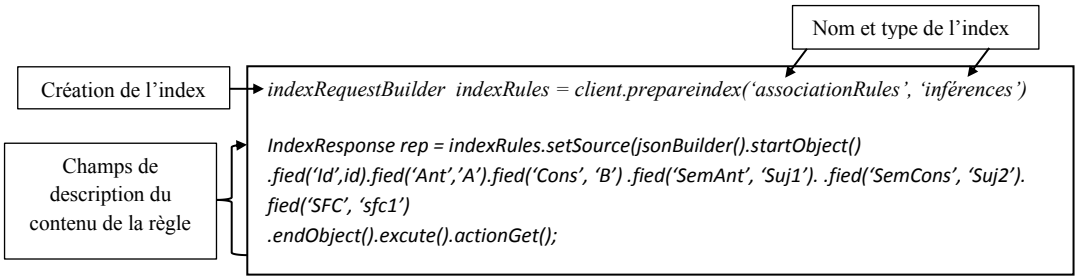
- Améliorer les corrélations entre les intérêts des utilisateurs lors de la sélection des règles d'inférence, notamment lorsque ces derniers partagent les mêmes sujets d'intérêt, mais n'emploient pas les mêmes étiquettes et requêtes (cf. Annexe 4). Ceci aide à améliorer le rappel du système, c'est-à-dire, sa capacité à repérer plus de documents pertinents pour les utilisateurs.
- Améliorer les données de recommandation en se basant sur un enrichissement sémantique à base de sujets d'intérêt. Ceci aide aussi à améliorer le rappel du système.

- Améliorer la précision du système en désambiguïsant le contenu des objets polysémique au sein des règles d'association. Ceci aide à éliminer les documents non pertinents.
- Faciliter et optimiser l'accès aux règles d'association en exploitant un index inversé pour la représentation de leur contenu et en adoptant une technique de sélection à deux niveaux qui combine la sélection générique par sujets d'intérêt et la sélection spécifique par données d'intérêt granulaires. Une technique de combinaison entre ces deux niveaux de sélection est alors proposée.

Nous commençons par présenter, dans la section qui suit, comment les règles d'association extraites sont stockées dans un index pour une meilleure exploitation, puis nous irons à l'évaluation du système exploitant ces règles au sein d'un processus d'inférence de données.

A. Indexation hybride des règles d'association

Nous avons vu dans les précédentes étapes (section VII.5.2.1), comment les données génériques des utilisateurs, notamment les sujets et les groupes de sujets, sont exploitées pour supprimer les calculs inutiles durant l'extraction des itemsets spécifiques fréquents. Nous avons pu constater aussi que dans certaines situations, les opérations supplémentaires effectuées sur les données génériques des utilisateurs peuvent avoir un impact négatif sur le temps de calcul de ces itemsets fréquents. Nous allons voir dans cette section, comment le traitement de ces données génériques peut avoir un impact positif sur l'amélioration de la pertinence des résultats. Il consiste à décrire sémantiquement le contenu de ces règles d'association avec les classes contextuelle auxquelles elles appartiennent et avec les sujets d'intérêt auxquels sont liés les objets antécédents et conséquents de ces règles. Pour ce faire, nous exploitons la structure d'index inversé. De la même manière qu'avec les documents, les règles sont structurées en champs de descriptions. Chaque règle est décrite avec un identifiant, l'ensemble de ses antécédents spécifiques et génériques, l'ensemble de ses conséquents spécifiques et génériques, ainsi que la classe contextuelle SFC_i à laquelle appartient cette règle. A cet effet, l'API d'indexation d'Elastic est exploitée. Voici un extrait du code de cette structure d'indexation.



Cette indexation par champs de description permet d'effectuer des sélections génériques et spécifiques selon le besoin. Nous avons parlé dans la section IV.5.4.4 du modèle théorique, sur les avantages que cet enrichissement peut avoir sur l'inférence de données. Nous évaluons alors la capacité de ce processus d'inférence à atteindre les objectifs visés dont le but principal est d'améliorer l'efficience du système (la pertinence des résultats et le temps de réponse). Ces objectifs se résument par les points suivants:

- Surmonter le problème de l'ambigüité des données polysémiques au sein des règles d'inférence.
- Améliorer la corrélation entre les intérêts des utilisateurs lors de la sélection des règles d'inférence et lors de l'inférence de données en question.
- Optimiser le temps d'accès aux règles d'inférence.

B. Évaluer la capacité du système à surmonter le problème d'ambiguïté des données au sein des règles d'inférence

Cette évaluation consiste à estimer la capacité du système à éliminer les documents de recommandation non pertinents lorsque les règles spécifiques exploitées contiennent un ou plusieurs objets ambigus. L'ambiguïté d'un objet (étiquette ou requête) peut être validée en vérifiant si cet objet est lié à plusieurs sujets dans les profils des utilisateurs. D'habitude, l'efficacité d'un système de recommandation est testée en évaluant la proportion des recommandations qui sont jugées pertinentes par les utilisateurs. Puisque dans notre cas nous avons affaire à des utilisateurs inconnus de la communauté Delicious, nous avons procédé de cette façon :

- Sélection d'un ensemble d'objets ambigus qui existent dans les règles d'association de notre système.
- Sélection des utilisateurs ayant exploité un ou plusieurs objets ambigus de la liste d'objets sélectionnés.

- Supprimer du profil de chaque utilisateur sélectionné, des documents qui sont liés à un ou plusieurs objets ambigus de la liste sélectionnée, puis appliquer notre approche sur le reste des données afin de montrer si le système arrive ou non à proposer les documents supprimés à leurs utilisateurs correspondants. Notons que tous les documents supprimés sont choisis au hasard afin de préserver l'intégrité de notre évaluation. Afin d'effectuer une validation croisée, pour chaque objet ambigu, l'ensemble des documents qui lui sont associés est divisé en plusieurs sous-ensembles, puis un sous-ensemble à supprimer est sélectionné dans chaque évaluation afin de l'utiliser comme ensemble de tests.
- Enfin, tester la capacité du système à localiser les documents supprimés. Cette capacité est évaluée en se basant sur les métriques de rappel et de précision.

$$Précision = \frac{Nombre\ de\ documents\ pertinents}{Nombre\ total\ de\ documents\ selectionnés}$$

$$Rappel = \frac{Nombre\ de\ documents\ pertinents}{Nombre\ total\ de\ documents\ \ dans\ l'ensemble\ de\ test}$$

Définition de la pertinence des documents de recommandation. Il est important de mettre l'accent sur la pertinence de nos résultats. Dans les systèmes qui se basent sur les cotes d'évaluations, un document est considéré comme pertinent si la différence entre la cote d'évaluation fournie par l'utilisateur et celle estimée par le système n'est pas grande. Dans un système n'utilisant pas les cotes d'évaluations comme le nôtre, un document pertinent est celui qui est susceptible d'être consulté par l'utilisateur. Étant donné que l'évaluation est fondée sur des utilisateurs que nous ne connaissons pas, le protocole d'évaluation est donc basé sur la suppression des documents de leurs profils, puis l'évaluation de la capacité du système à récupérer ces documents. Ainsi, un document est considéré comme pertinent s'il appartient à la liste des documents supprimés. Le problème avec cette pertinence est que notre système peut proposer d'autres documents qui peuvent être pertinents, mais n'appartiennent pas aux documents supprimés, car notre système se base sur l'extraction de corrélations entre les utilisateurs à base de sujets d'intérêts. C'est la raison pour laquelle nous considérons différentes pertinences selon chaque tâche d'évaluation.

Dans cette section, la tâche d'évaluation consiste à tester la capacité du système à lever l'ambigüité sur les données polysémiques, ceci se traduit par le fait que ce système arrive à éliminer les

documents qui ne sont pas liés aux sujets d'intérêt de l'utilisateur. Ainsi, un document est considéré comme pertinent s'il appartient à l'ensemble des documents de tests ou appartient aux mêmes sujets d'intérêt auxquels ces documents tests appartiennent.

Évaluation et comparaison. Nous comparons les résultats obtenus par notre système avec ceux obtenus avec les travaux de (Beldjoudi *et al.* 2017). Ces auteurs adoptent, pour sélectionner les documents pertinents, la similarité entre les utilisateurs à base de leurs relations sociales, notamment leurs annotations. Nous avons soulevé dans la section IV.4 (cf. page 135), les faiblesses que cette méthode peut avoir lorsqu'ils existent des utilisateurs qui s'intéressent à la fois à différents sujets auxquels est lié l'objet ambigu. Pour valider ces faiblesses, nous augmentons explicitement les profils de certains utilisateurs de test avec des documents qui couvrent d'autres sujets auxquels sont liés les objets ambigus sélectionnés. Par exemple, les profils ayant des documents sur les virus informatiques sont enrichis avec des documents qui couvrent le sujet des virus humains, puis nous évaluons la capacité du modèle à écarter ces documents lors de la recommandation des documents.

Les résultats de la figure 7.20 montrent que notre modèle est meilleur en précision et en rappel avec respectivement 39% et 81% d'amélioration par rapport au modèle de référence (Beldjoudi *et al.* 2017). Les résultats obtenus avec ce modèle de référence montrent que la technique proposée renvoie les documents qui ont été ajoutés explicitement. Ceci confirme que dans de telles situations, le calcul des relations sociales entre les utilisateurs présente une faiblesse. Ceci explique la baisse de précision de leur système.

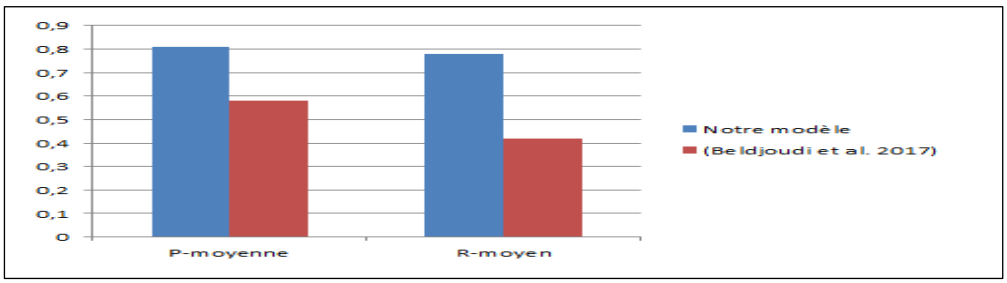


Figure 7. 20. Valeurs de précision moyenne, rappel moyen et de F1 de la recommandation de données de notre système et du système Baseline

De plus, ces auteurs n'ont pas traité la polysémie des objets lorsqu'ils se trouvent dans l'antécédent de la règle (par exemple Java → tools). Autrement dit, ce problème d'ambigüité a été traité uniquement durant l'inférence de données c'est-à-dire, lorsque les objets polysémiques sont présents dans la conséquence de

la règle, et a été négligé lors de la sélection des règles. De cette façon, la sélection d'une règle d'inférence se base sur une simple correspondance syntaxique entre les objets antécédents de cette règle et les données d'intérêt de l'utilisateur dans son profil (Beldjoudi *et al.* 2017). De cette façon, lorsque de telles règles d'inférence « Java → tools » et « Java→ open source » existent, tous les utilisateurs qui ont utilisé le mot Java pour annoter/chercher des documents recevront des documents qui sont liés à « tools » et à « open source » car le système ne pourra déduire de quel Java l'utilisateur est intéressé. Tandis qu'avec notre modèle, d'un côté les intérêts des utilisateurs sont décrits sémantiquement avec des sujets d'intérêt, et d'un autre, le système exploite la base d'indexation où les règles sont décrites aussi sémantiquement avec des sujets d'intérêt. De cette façon, les deux règles de l'exemple précité ne sont exploitées que pour un utilisateur qui s'intéresse à Java appartenant aux mêmes sujets qui décrivent ces règles dans l'index.

En ce qui concerne le rappel du système, les auteurs dans (Beldjoudi *et al.* 2017) utilisent uniquement des règles spécifiques construites à partir des étiquettes d'annotation des utilisateurs. Ceci limite la détection des données pertinentes pour la recommandation. Cela est dû à différentes raisons:

- Lorsque les utilisateurs partagent des sujets communs, mais expriment différemment leurs besoins en information à travers les étiquettes d'annotations. Ceci limite la détection de corrélations entre leurs intérêts lors de la section des règles d'inférence et lors de l'inférence de données en question. Plus concrètement :
    - Lorsque la sélection des règles est basée uniquement sur une correspondance exacte entre les objets antécédents des règles et les intérêts des utilisateurs, ceci limite la sélection d'autres règles pouvant être utiles pour la recommandation de données pertinentes.
    - Lorsque les relations sémantiques entre les données d'intérêt spécifiques des utilisateurs ne sont pas identifiées, ceci limite les techniques qui peuvent être appliquées lors de l'inférence de données pour proposer les documents candidats à la recommandation. Le système se limite à proposer uniquement les documents qui sont liés aux objets conséquents des règles sélectionnées.

Ainsi, le recours à un enrichissement sémantique dans de tels cas aide à augmenter le rappel du système.

Afin de valider l'efficacité de cet enrichissement sur l'amélioration du système, nous soutenons les résultats obtenus dans cette section, avec d'autres expérimentations que nous présentons dans la section suivante où un groupe particulier d'utilisateurs est exploité dans l'évaluation de cette technique

d'enrichissement (cf. section C). Il s'agit des utilisateurs ayant des données d'intérêt qui présentent uniquement des corrélations génériques avec les intérêts fréquents qui constituent les règles d'inférence du système.

C. Évaluer l'efficacité du système à augmenter le rappel du système en améliorant la corrélation entre les intérêts des utilisateurs

Cette section a pour objectif d'évaluer la capacité du système à améliorer l'extraction de corrélations entre les intérêts des utilisateurs durant le processus de recommandation de données en vue d'améliorer le rappel du système. Cette amélioration est possible lors de :

i) la sélection des règles d'association correspondant aux intérêts des utilisateurs en appliquant une technique de description et de sélection hybride de leur contenu,

ii) et lors de l'inférence de données en enrichissant la liste des recommandations avec d'autres éléments connexes à base de sujets d'intérêt.

Pour ce faire, le processus de recommandation est testé sur un groupe d'utilisateurs dont les intérêts présentent uniquement des corrélations génériques avec les intérêts fréquents des autres utilisateurs système. Il consiste à tester si le système est capable de recommander des documents pour les utilisateurs qui n'ont pas de corrélations spécifiques avec les autres utilisateurs. On se référant à la figure 7.19 qui résume le protocole global d'évaluation du processus d'inférence, il s'agit d'exploiter le groupe nommé par « groupe 1 d'utilisateurs ». Nous pouvons voir que cet ensemble d'utilisateurs est lié aux itemsets spécifiques non fréquents qui sont éliminés de la phase d'extraction des itemsets fréquents. Ces itemsets sont à leur tour liés à des itemsets génériques fréquents. Ces utilisateurs sont donc adaptés à cette tâche d'évaluation. Cette dernière consiste à tester la capacité de la technique d'indexation hybride des règles à améliorer la localisation des règles qui sont pertinentes pour les utilisateurs. Une règle est considérée comme pertinente pour un utilisateur donné lorsque son exploitation permet de lui proposer des documents pertinents. Dans cette étape d'évaluation, un document est considéré comme pertinent pour un utilisateur s'il appartient à la liste des documents consultés par cet utilisateur. Il consiste alors à les supprimer de son profil puis tester si le système est capable de les récupérer. Pour ce faire, le protocole d'évaluation suivant est appliqué :

1. Nous sélectionnons aléatoirement un sous-ensemble d'objets qui constituent les règles d'association génériques du système. C'est-à-dire, un sous-ensemble de sujets d'intérêt fréquents des utilisateurs.
2. Préparation d'un groupe d'utilisateurs ayant des intérêts spécifiques non fréquents qui appartiennent aux sujets d'intérêt fréquents sélectionnés dans l'étape 1.
3. Pour chaque utilisateur dans le groupe préparé, nous sélectionnons un sous-ensemble de documents à supprimer de son profil. Il consiste à sélectionner d'abord les documents qui sont liés aux sujets fréquents sélectionnés, puis sélectionner aléatoirement un sous-ensemble à supprimer. Ce sous-ensemble de documents est considéré comme l'ensemble de tests.
4. Appliquer notre approche d'extraction de règles sur le reste des données des utilisateurs. Cela permet d'extraire les corrélations entre les intérêts des utilisateurs système sur un sous-ensemble de données d'apprentissage et garder le sous-ensemble restant pour les tests.
5. Puis, tester si le système arrive ou non à proposer les documents supprimés à leurs utilisateurs correspondants. Dans cette étape, deux techniques sont appliquées pour la sélection des règles d'inférence. La première est fondée sur une recherche spécifique des règles, que nous appelons par la sélection monodimensionnelle, et la deuxième technique est une sélection à deux niveaux de sélection (générique puis spécifique), que nous appelons par la sélection hybride.
6. Nous évaluons la pertinence des résultats obtenus par les deux techniques et les comparons l'une à l'autre. Cette pertinence est évaluée par estimation de la précision et du rappel du système définis à travers les deux métriques ci-dessus. Un document est considéré comme pertinent s'il appartient à la liste des documents supprimés. Ce protocole d'évaluation est résumé dans la figure 7.21.

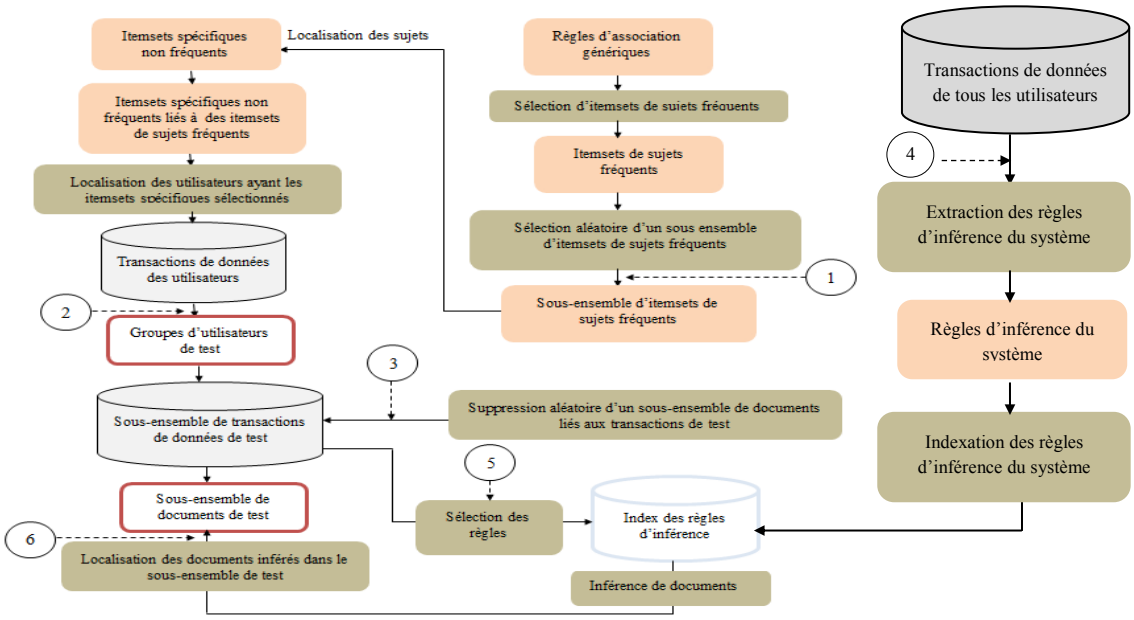


Figure 7. 21. Protocole d'évaluation de l'efficacité de la technique à deux niveaux de sélection des règles dans le système d'inférence de données d'intérêt des utilisateurs

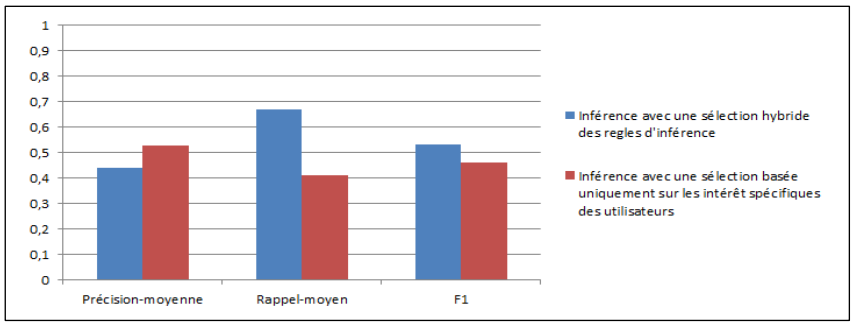


Figure 7. 22. La pertinence des résultats obtenus du système de recommandation suite à deux techniques de sélection des règles d'inférence

Les résultats de la figure 7.22 montrent que par rapport à la technique de sélection monodimensionnelle, la technique de sélection hybride donne un meilleur rappel au système, mais dégrade la précision des résultats. Par ailleurs, cette technique de sélection hybride donne un meilleur compromis évalué à travers la métrique F1.

En ce qui concerne la technique de recommandation par enrichissement de données d'inférence, celle-ci consiste à élargir la liste de recommandations avec d'autres documents qui soient liés aux mêmes sujets de recherche que les documents qui sont proposés depuis les conséquents des règles d'inférence. Les résultats de la figure 7.23, montrent que lorsque la pertinence d'un document de recommandation est

validée uniquement lorsqu'il appartient à l'ensemble de documents déjà consultés par l'utilisateur (pertinence spécifique), c'est-à-dire, l'ensemble des documents de tests qui sont supprimés de son profil, cela donne une précision faible (cf. figure 23 partie 1) par rapport au cas où cette pertinence est définie comme étant un document qui soit lié à un sujet d'intérêt de cet utilisateur (pertinence générique), c'est-à-dire, à un des sujets auxquels sont liés les documents de tests (cf. figure 7.23 partie 2). Cette faiblesse en précision est justifiée par le fait que lorsque système se base sur une pertinence à base de sujets d'intérêt pour sélectionner les documents, il renvoie aussi d'autres documents connexes. Puisque nous avons affaire à des utilisateurs que nous ne connaissons pas, nous ne pouvons donc pas savoir si tous les résultats connexes qui sont proposés à ces utilisateurs peuvent être ou non intéressants pour lui. Ainsi, nous estimons la performance de notre processus uniquement à travers la métrique de rappel qui estime sa capacité à sélectionner tous les documents qui sont supprimés de son profil. Les résultats illustrés dans la partie 1 de la figure 7.23 montrent que le système obtient un meilleur rappel lorsque la technique d'enrichissement est appliquée lors de l'inférence de données. Comme nous pouvons le voir, la faiblesse en précision engendre une baisse en compromis F1 par rapport à la technique d'inférence où l'enrichissement de données n'est pas considéré.

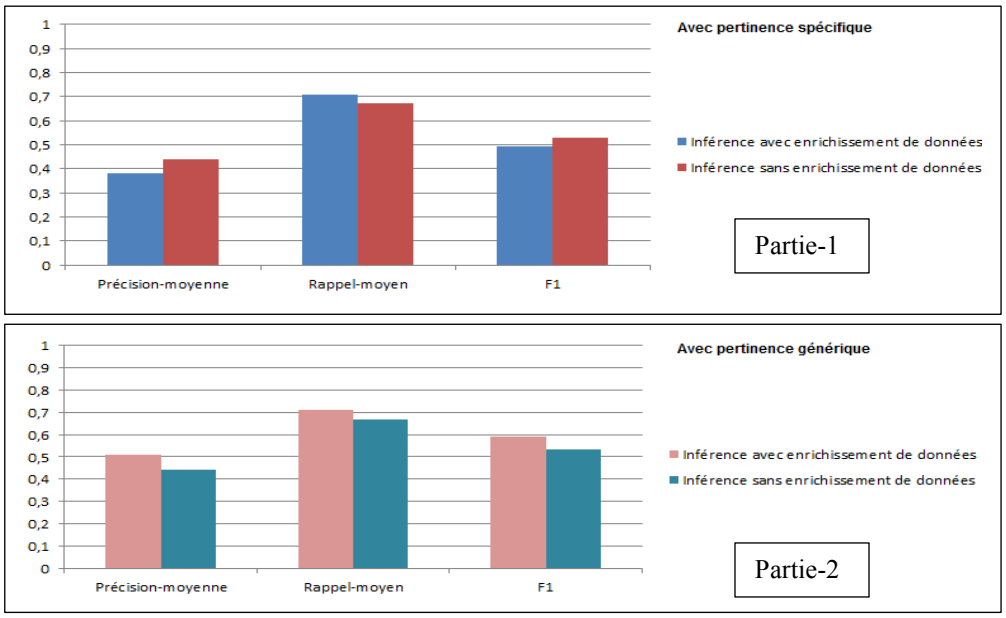


Figure 7. 23. Comparaison de la pertinence des résultats qui sont obtenus des deux techniques d'inférence d'intérêts: inférence avec et sans enrichissement de données à base de sujets d'intérêt

	Définition de la pertinence	Liste de documents de tests	Métriques d'évaluation
Pertinence spécifique	Un document est dit pertinent pour un utilisateur s'il a été consulté par cet utilisateur	Les documents de tests sont les documents qui sont supprimés depuis un profil utilisateur.	Précision = $\frac{nombre\ de\ douments\ pertinents\ selectionnés}{nombre\ de\ tous\ les\ documents\ selectionnés}$ Rappel = $\frac{nombre\ de\ douments\ pertinents\ selectionnés}{nombre\ de\ documents\ dans\ la\ liste\ de\ tests}$ F1= $\frac{2*précision*rappel}{Précision+rappel}$
Pertinence générique	Un document est dit pertinent pour un utilisateur s'il a été consulté par cet utilisateur ou appartient au même sujet de recherche qu'un des documents consultés par lui.	Les documents de tests sont les documents qui sont supprimés depuis le profil utilisateur ainsi que les autres documents qui appartiennent aux mêmes sujets de recherche que les documents supprimés.	

D. Évaluation de l'impact de la similarité entre utilisateurs sur la qualité des résultats de recommandation

Nous avons vu dans la section précédente que lorsque le système se base sur les sujets d'intérêt des utilisateurs pour sélectionner les règles d'association intéressantes (le cas des utilisateurs n'ayant pas de corrélations spécifiques avec les autres utilisateurs) ou pour étendre la liste des documents candidats à la recommandation avec d'autres éléments connexes, cela donne un meilleur rappel au système par rapport au cas où ces deux techniques ne sont pas appliquées, mais il baisse par contre sa précision. Dans cette section, nous allons voir si la précision de ces résultats de recommandation peut être améliorée en renforçant le critère de sélection des règles d'inférence et des documents candidats à la recommandation, par l'intégration de la similarité entre les utilisateurs. Cela se traduit comme suit :

Lors de la sélection des règles d'inférence : lorsqu'une règle d'inférence R_i ne corresponds pas aux intérêts spécifiques de l'utilisateur cible u et couvre ses sujets d'intérêts, le système n'exploite cette règle pour recommander des données à cet utilisateur que lorsque les objets antécédents de R_i ont été exploités aussi par les utilisateurs voisins de cet utilisateur. Ces voisins représentent l'ensemble des utilisateurs ayant une forte similarité avec l'utilisateur cible (cf. section IV.5.4.4.1).

Lors de la sélection des documents candidats à la recommandation : lorsque la liste des recommandations est enrichie avec d'autres documents connexes à base de sujets d'intérêt, le système propose parmi cette liste d'enrichissement uniquement les documents qui ont été consultés par les utilisateurs similaires.

Nous étudions l'impact de cette similarité sur la performance du système. Cette performance est évaluée à travers le compromis F1. Elle dépend du seuil de similarité que le système définit pour considérer deux utilisateurs similaires. Le seuil optimal est celui qui permet de donner une valeur de F1 maximum au système. Les expérimentations ont pu fixer un seuil de similarité de 0.4. Les résultats obtenus avec ce seuil sont illustrés dans la figure 7.24.

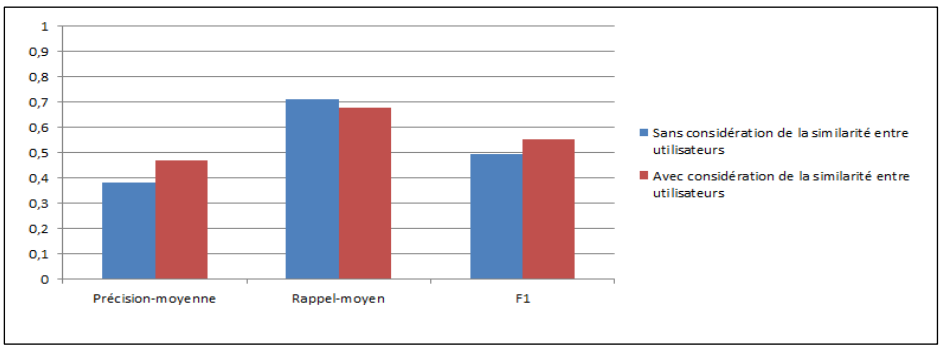


Figure 7. 24. Impact de la similarité entre utilisateurs sur la qualité des résultats de recommandation

Les résultats de la figure 7.24 montrent que lorsque la similarité entre les utilisateurs est prise en considération à la fois dans la sélection des règles d'inférence et dans l'enrichissement de la liste des documents candidats à la recommandation, cela améliore la précision du système et baisse légèrement son rappel, mais donne une amélioration de 12% dans la performance globale du système évaluée à travers le compromis F1 (précision-rappel).

E. Évaluation de l'efficacité du système à optimiser l'accès aux règles d'inférence

Cette section évalue la capacité du système à optimiser l'accès aux règles d'inférence lorsque :

- Un index inversé est exploité pour décrire ces règles,
- Et lorsqu'une description hybride (spécifique et générique) de leur contenu est adoptée au sein de cet index inversé.

Il consiste à évaluer le temps moyen nécessaire pour localiser les règles intéressantes pour les utilisateurs. Une règle est considérée comme intéressante pour un utilisateur donné lorsque les objets antécédents de cette règle couvrent les intérêts de cet utilisateur.

Dans les deux taches d'évaluations, l'impact de deux facteurs est étudié sur le temps de sélection des règles. A savoir, le nombre des règles d'inférence du système et le nombre moyen des données d'intérêts des utilisateurs qui sont prises en considération durant la sélection des règles.

Afin de démontrer l'efficacité de l'index inversé à réduire le temps d'accès aux règles d'inférence, nous le comparons à une structure classique où les règles sont sauvegardées dans un fichier de données et l'accès est basé sur la lecture du contenu du fichier.

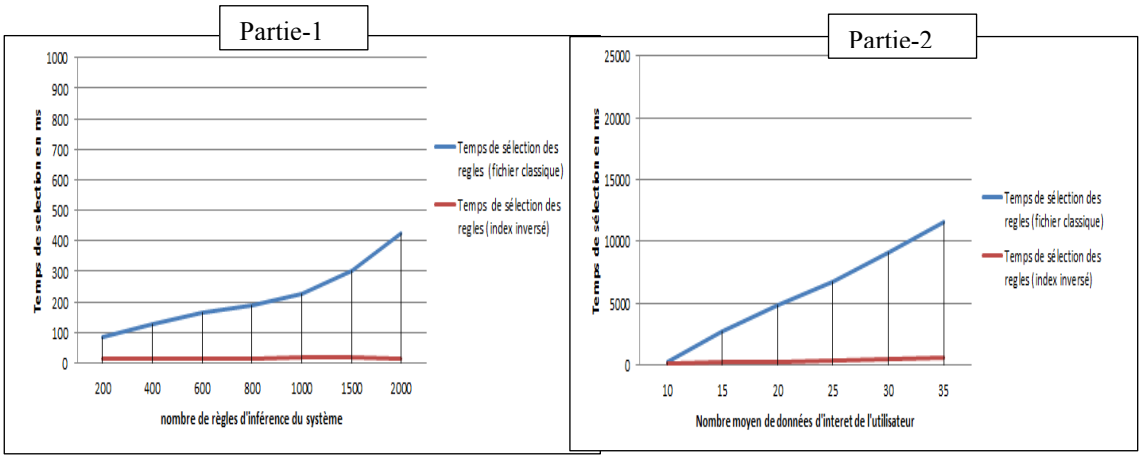


Figure 7. 25. Évaluation de l'efficacité d'une représentation des règles à base d'index inversé : étude de l'impact du nombre de règles système et du nombre de données d'intérêt des utilisateurs sur le temps d'accès aux règles d'inférence

Les résultats de la partie 1 de la figure 7.25 montrent que lorsqu'un fichier de données est adopté pour stocker les règles du système, les deux facteurs étudiés ont un impact plus marqué sur le temps de sélection des règles par rapport à l'index inversé. Plus concrètement, on remarque qu'un pas d'incrémentation de 1000 règles engendre une augmentation de 86% en temps d'accès lorsqu'un fichier de données classique est utilisé pour la représentation des règles. Tandis qu'avec l'index inversé ce pas d'incrémentation n'a presque pas d'impact sur le temps d'accès. La même chose est observable avec le nombre moyen des données d'intérêt des utilisateurs qui sont prises en considération lors de la sélection des règles (cf. figure 7.25 partie 2), on remarque que le temps de sélection des règles avec un fichier classique est de 11524 ms lorsqu'il y a 35 données d'intérêt à considérer durant la sélection, il est énormément supérieur au temps de sélection nécessaire à travers un index inversé qui est de 611 ms. En outre, on remarque que l'augmentation et l'évolution de données d'intérêt engendrent une dégradation considérable en temps d'accès par rapport à l'index inversé.

Afin de tester l'efficacité de la technique de description hybride des règles d'inférence, à réduire le temps de leur sélection, deux techniques de sélection de ces règles sont appliquées et comparées l'une à l'autre. La première s'appuie sur la sélection des règles d'inférence à travers un seul niveau spécifique de description. Il consiste à effectuer une recherche personnalisée sur le champ de description spécifique des règles. La deuxième s'appuie une sélection hybride à travers deux niveaux de sélection. Il consiste à sélectionner d'abord les règles qui couvrent les sujets d'intérêt de l'utilisateur puis filtrer parmi les

résultats obtenus celles qui correspondent à ses intérêts spécifiques. Puis, le temps d'accès aux règles intéressantes pour les utilisateurs, est évalué.

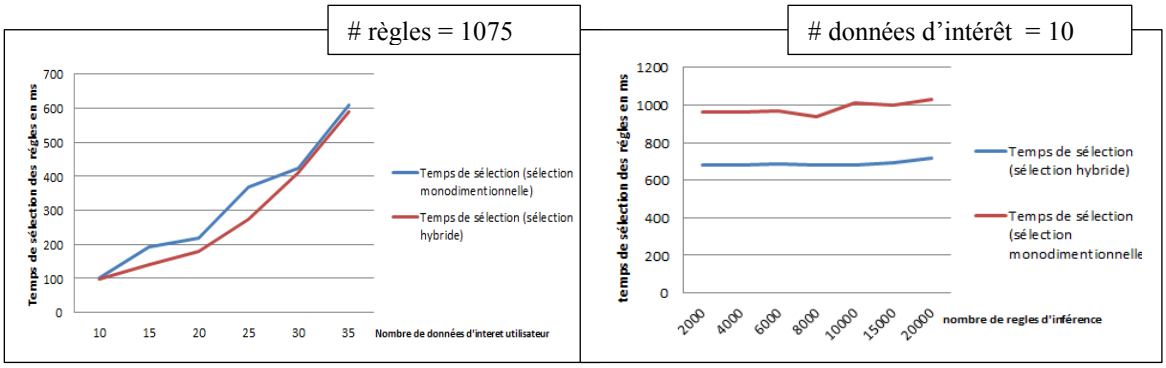


Figure 7. 26. Comparaison de l'efficience de deux techniques de sélection des règles au sein d'un index inversé: la sélection monodimensionnelle et la sélection hybride

Les résultats de la figure 7.26, montrent que la sélection hybride des règles d'inférence est meilleure qu'une sélection monodimensionnelle. Cette optimisation en temps de sélection s'explique par le fait qu'avec une sélection hybride, le système se base en premier lieu sur un accès personnalisé à base de sujets d'intérêt. Ceci permet d'écarter les règles qui ne couvrent pas les sujets d'intérêt de l'utilisateur, puis filtre parmi les règles résultantes celles qui correspondent à ses intérêts spécifiques. Le premier filtre générique peut éliminer une grande quantité de règles non pertinentes à travers quelques accès, ce qui permet d'optimiser plusieurs accès spécifiques à l'index.

À partir des résultats des deux figures 7.25 et 7.26 nous pouvons valider l'efficacité de l'index inversé ainsi que la description hybride choisie pour décrire les règles d'inférence dans l'amélioration de la performance du processus d'inférence d'intérêts des utilisateurs.

VII.5.2.4. Évaluation de la personnalisation du processus d'inférence de données

Cette personnalisation a pour objectif d'améliorer la prédiction des intérêts de l'utilisateur en exploitant depuis son profil uniquement les données les plus représentatives de ses intérêts courants, c'est-à-dire, ses préférences. A cet effet, une fonction de pertinence multidimensionnelle est proposée pour calculer les préférences de l'utilisateur qui permettent la sélection des règles d'association les plus appropriées pour la prédiction de ses futurs intérêts. Cette pertinence combine l'aspect temporel et fréquentiel des données dans le profil de l'utilisateur au sein d'une fonction pondérée (cf. section IV.5.4.4.1). Cette combinaison est basée sur l'hypothèse suivante : « les besoins de l'utilisateur évoluent

au fil du temps et certaines données enregistrées dans son profil peuvent devenir obsolètes et non pertinentes, il est donc important de prendre en considération la fraicheur de ces intérêts. Dans d'autres cas, les données d'intérêt récentes de l'utilisateur peuvent être liées à un besoin spécifique et temporaire qui ne représente pas ses centres d'intérêt récurrents, il est donc important de prendre en considération aussi la fréquence des données d'intérêts».

Afin de tester l'efficacité de cette pertinence multidimensionnelle sur le calcul des préférences des utilisateurs et la prédiction de leurs futurs intérêts, le profil de chaque utilisateur est délimité en deux sous-ensembles de données. Le premier est constitué d'un ensemble de (k-n) activités utilisées pour apprendre ses intérêts, le deuxième est le reste des n-activités qui sont utilisées pour tester la pertinence des prédictions du système. Pour cela, les activités de l'utilisateur sont ordonnées chronologiquement selon leur fraicheur puis un point de coupure est appliqué pour délimiter les deux sous-ensembles d'activités. Les règles d'association sont extraites uniquement depuis les activités d'apprentissage des utilisateurs et sont exploitées pour l'inférence de leurs intérêts. Nous estimons ensuite la capacité du système à prédire les activités de test de chaque utilisateur apparaissant après la coupure. Dans cette étape d'évaluation, les données prédites ne représentent pas des documents, mais des activités de recherche, c'est-à-dire, des données spécifiques (requêtes/ étiquettes).

Afin de tester l'influence des deux aspects (fréquentiel et temporel) sur la qualité des prédictions, deux types de coupure sont appliqués sur les activités des utilisateurs. Le premier cas consiste à appliquer une coupure entre deux activités temporellement proches (une semaine ou moins), cela permet de simuler un comportement actif de l'utilisateur. Le deuxième cas consiste à appliquer cette coupure entre deux activités éloignées (un mois et plus), cela permet de simuler une période d'inactivité de l'utilisateur. Puis tester la capacité du système à prédire les prochaines activités qui apparaissent après la coupure. Cela permet d'étudier l'impact des deux facteurs étudiés (notamment le facteur temporel et le facteur fréquentiel) sur la qualité des prédictions lorsque différentes périodes de temps séparent l'instant présent de la prédiction de l'ensemble des activités récentes des utilisateurs.

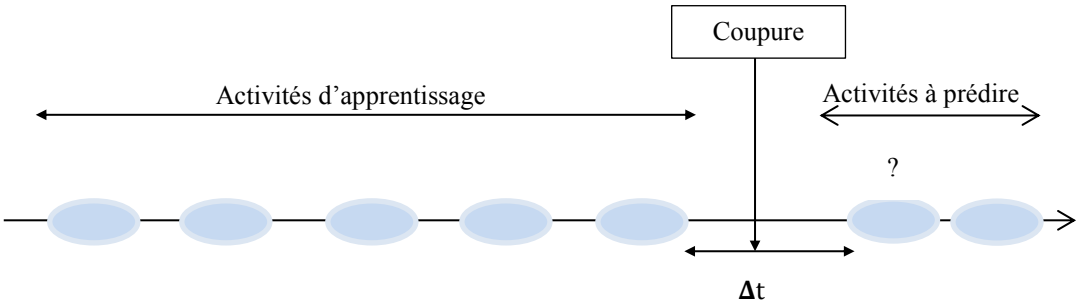


Figure 7. 27. Protocole d'évaluation de la prédiction personnalisée des intérêts de l'utilisateur

Pour chaque type de coupure, nous faisons varier de 0 à 1 la valeur du coefficient α. Ce coefficient définit le degré d'importance des deux facteurs étudiés. Plusieurs cas sont alors considérés :

- Lorsque α = 0 : seul l'aspect fréquentiel est pris en considération dans le calcul des préférences de l'utilisateur.
- Lorsque α ∈ [0.1, 0.4] : l'aspect fréquentiel a plus d'importance dans le calcul des préférences utilisateur.
- Lorsque α = 0.5 : les deux aspects ont le même degré d'importance.
- Lorsque α ∈ [0.6, 0.9] : l'aspect temporel a plus d'importance que l'aspect fréquentiel.
- Lorsque α = 1 : seul l'aspect temporel est pris en considération dans le calcul des préférences utilisateur.

A chaque valeur d'alpha, la précision du système est évaluée comme suit :

$$Précision = \frac{Nombre\ d'activités\ pertinentes}{Nombre\ total\ des\ activités\ prédites}$$

La valeur de précision est calculée pour évaluer les prédictions de chaque utilisateur puis la moyenne est calculée pour tous les utilisateurs. Une activité est considérée comme pertinente pour un utilisateur si elle appartient à l'ensemble de ses activités de test ou appartient au même sujet d'intérêt auquel sont liées ces activités.

Pour résumer, l'évaluation de la qualité de la prédiction personnalisée est effectuée en étudiant l'impact de :

- La période **Δ**t séparant l'instant de prédiction des autres activités récentes de l'utilisateur,

- Et du degré d'importance des deux facteurs temporel et fréquentiel qui sont pris en considération dans le calcul des préférences de l'utilisateur.

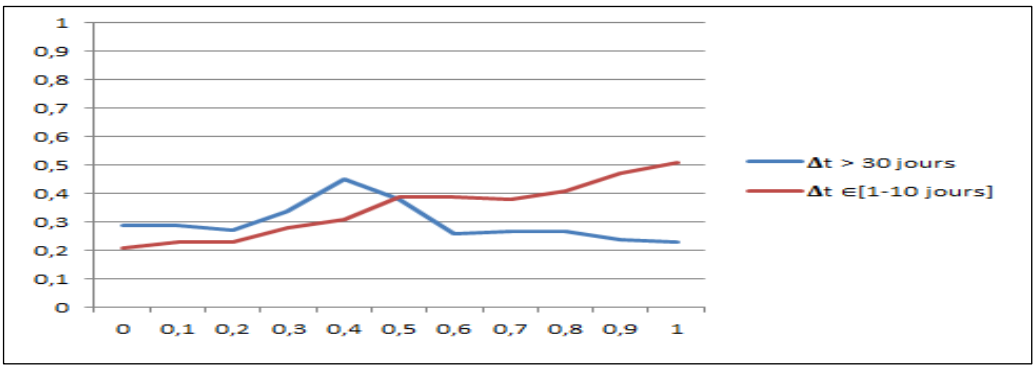


Figure 7. 28. Étude de l'impact des deux critères de fréquence et de fraicheur des données d'intérêt sur le calcul des préférences de l'utilisateur et la prédiction de ses intérêts

Les résultats de la figure 7.28 montrent que :

Cas 1 : lorsque la période **Δ**t est petite (**Δ**t < 10 jours), la pertinence des résultats est meilleure lorsque le facteur temporel a plus d'impact ($\alpha \in [0.6, 1]$). Cela indique que dans un tel cas, les données les plus représentatives des intérêts courants de l'utilisateur sont liées aux données les plus fraiches dans son profil.

Cas 2 : dans le cas où **Δ**t est grande (**Δ**t> 30 jours), on remarque que :

- L'intervalle $\alpha \in [0.3, 0.5]$ a donné de meilleures valeurs de précision en ce qui concerne la prédiction des intérêts des utilisateurs. Cela indique que lorsqu'il y a une grande période de temps entre l'instant de recommandation et les activités récentes de l'utilisateur (ceci peut être traduit en réalité par une grande période d'inactivité de l'utilisateur puis un retour sur le système), le facteur fréquentiel a plus d'importance que le facteur temporel sur l'extraction des préférences de cet utilisateur. On remarque aussi que la précision diminue à partir de $\alpha = 0.6$ (lorsque le facteur temporel a plus d'importance), cela indique que les intérêts les plus frais dans le profil de l'utilisateur ne sont pas pertinents lorsqu'ils ne sont pas fréquents.
- Lorsque $\alpha \in [0, 0.3]$, c'est-à-dire, lorsque soit l'aspect temporel n'est pas pris en considération ($\alpha = 0$), ou lorsque l'aspect temporel a une importance faible ($\alpha \in [0.1, 0.3]$), on remarque que la précision est faible. Cela indique que l'aspect temporel a également une importance dans le calcul

des préférences de l'utilisateur et que les intérêts fréquents dans le profil de l'utilisateur n'ont pas d'importance si leur fraicheur est faible.

Ce que nous pouvons conclure est que dans ce deuxième cas (**Δ**t> 30 jours), la sélection des données pertinentes dans le profil de l'utilisateur représentant ses préférences se base sur un compromis entre les deux facteurs fréquentiel et temporel. On remarque que la valeur de $\alpha = 0.4$ a donné la meilleure précision au système. Cela donne un degré d'importance de 40% au facteur temporel et un degré d'importance de 60% au facteur fréquentiel de données. Ainsi nous pouvons conclure qu'en cas d'inactivité de l'utilisateur sur le système, lors de son retour sur le système, il est plus susceptible de consulter des données qui sont liées aux intérêts récents les plus fréquents dans son profil. Ainsi, lorsque les différents intérêts de l'utilisateur ont des valeurs de fréquence qui convergent, le critère de fraicheur est celui qui départage entre eux pour évaluer les préférences de l'utilisateur qui servent à prédire ses futurs intérêts.

Il arrive que les données d'intérêt les plus fraîches représentent en même temps les données les plus fréquentes chez l'utilisateur et vice versa, cela justifie :

- Une valeur de précision satisfaisante lorsque les deux facteurs ont la même importance ($\alpha = 0.5$) dans les deux cas étudiés (**Δ**t ∈ [1-10 jours] et **Δ**t> 30 jours).
- La convergence des résultats de précision dans les deux cas étudiés de **Δ**t, lorsque l'un des deux facteurs a plus d'importance que l'autre, c'est-à-dire, lorsque $\alpha \in [0\text{-}0.4]$ et lorsque $\alpha \in [0.6\text{-}1]$.

Jusqu'à présent, nous avons présenté le protocole de construction du profil utilisateur et l'efficacité de sa structure multidimensionnelle dans l'inférence de données proposée pour enrichir davantage son contenu. Nous présentons dans la section prochaine comment le contenu de ce profil peut être aussi intégré pour personnaliser les résultats des utilisateurs lorsque ces derniers expriment explicitement leurs besoins en information à travers des requêtes de recherche.

VII.5.3. Évaluation du système de recherche d'information personnalisé (SRIP) par intégration du profil utilisateur

Ce modèle personnalise les résultats de recherche de l'utilisateur en fonction de ses centres d'intérêt qui sont recueillis suite à ses interactions avec le système et stockés dans son profil. Nous avons vu dans le chapitre 5, comment ces centres d'intérêt sont exploités dans nos études pour personnaliser

l'accès à l'information. Nous avons proposé deux modèles qui exploitent ces intérêts en deux façons différentes. Le premier modèle les exploite au niveau de l'indexation des documents et propose une représentation centrée-utilisateur de leur contenu. Le deuxième modèle les exploite dans le réordonnancement des résultats de recherche. Nous évaluons dans cette section l'efficacité de chacun de ces deux modèles en les comparants aux deux modèles de référence, notamment le modèle classique BM25F et le modèle social BM25FS, et en les comparants l'un à l'autre.

VII.5.3.1 Évaluation du modèle de l'indexation personnalisée des documents

Ce modèle représente une amélioration du modèle classique BM25F qui ne tient pas compte des centres d'intérêt de l'utilisateur lors de ses recherches d'information (Robertson *et al.* 2004), et du modèle social BM25FS qui propose d'améliorer les recherches de l'utilisateur en exploitant ces centres d'intérêt (notamment les étiquettes d'annotation de l'utilisateur) dans la description des documents (Bouhini 2014). Ce dernier ne tient pas en compte du contexte et de la représentativité des données durant la description des documents. Ces deux paramètres sont appelés dans nos études par les critères de pertinence des données d'enrichissement. Plusieurs faiblesses ont été soulevées de ces deux modèles (cf. section V.2.2.5), nous évaluons la validité de ces faiblesses en comparant les résultats obtenus de notre modèle aux résultats de ces deux modèles de référence. Cette évaluation comprend cinq volets d'étude:

- Lorsque les centres d'intérêt des utilisateurs sont pris en compte dans la description des documents.
- Lorsque les critères de pertinence des données d'enrichissement sont pris en compte durant la description des documents.
- Lorsque les intérêts du voisinage utilisateur sont également exploités dans la description personnalisée des documents en plus des intérêts individuels de l'utilisateur.
- Lorsque le voisinage de l'utilisateur est considéré comme une entité dynamique qui évolue au fil du temps.
- Lorsque la structure adoptée pour la description des documents est basée sur les champs de description qui aide à effectuer une recherche personnalisée. Cette recherche cible les k-champs de description qui sont liés aux intérêts de l'utilisateur et ceux de son voisinage (cf. section V.2.2.5 page 189).

A. Évaluation de l'impact d'une description orientée-utilisateur sur la RI

Afin de valider l'importance des intérêts utilisateur dans l'amélioration de ses résultats de recherche, en particulier lorsque ces données d'intérêt sont exploitées pour enrichir le contenu des documents interrogés, deux techniques d'indexation sont considérées :

- La première technique est l'indexation personnalisée qui intègre les intérêts des utilisateurs dans la description des documents (cf. section V.2.1).
- La deuxième technique est l'indexation classique mise en œuvre dans la section IV 2.2 qui décrit les documents avec leur contenu multidimensionnel sans tenir compte des intérêts de l'utilisateur.

Le protocole d'évaluation du processus de recherche d'information personnalisée (RIP) est effectué comme suit : les k-1 activités de chaque sujet intérêt de chaque utilisateur sont exploitées pour la description des documents et la $k^{ème}$ activité est exploitée pour tester la recherche personnalisée. Une comparaison est ensuite effectuée entre le modèle de recherche classique qui exploite l'index documentaire classique pour répondre aux requêtes des utilisateurs et le modèle personnalisé qui exploite l'index documentaire étendu avec les intérêts de ces utilisateurs pour répondre à ces requêtes. Nous présentons à travers le tableau 7.7 les résultats de cette comparaison en termes de pourcentages d'améliorations en rappel et en précision du modèle de la recherche personnalisée par rapport au modèle classique.

Avec intégration des intérêts utilisateurs	**P10**	**P20**	**MAP**	**Rappel-moyen**
Pourcentage d'amélioration	54%	22%	17%	29%

Tableau 7. 7. Pourcentage d'amélioration de la RI lorsque les intérêts des utilisateurs sont intégrés dans la description des documents

Les améliorations obtenues avec le SRI sont prometteuses lorsque les intérêts des utilisateurs sont pris en considération dans la description des documents. Cela est prévisible, car cette technique de description a permis d'augmenter le degré de correspondance des documents avec la requête, en particulier lorsque cette requête est liée à un besoin récurrent de l'utilisateur. Cela est fait en augmentant le nombre d'occurrences des jetons dans le contenu des documents lorsqu'ils correspondent aux intérêts de l'utilisateur, ce qui leur permet de gagner en classement dans la liste renvoyée. Ceci explique l'amélioration considérable en précisions P10 et en rappel.

Dans certains cas, l'enrichissement personnalisée des documents peut être inefficace et n'améliore pas la recherche de l'utilisateur, en particulier lorsque ces document sont enrichis avec des données qui ne correspondent pas à leur contenu. Il s'agit du cas où les données d'intérêt des utilisateurs sont liées à plusieurs interprétations que le système considère comme étant un seul intérêt. Le système doit donc détecter les données d'enrichissement adéquates au contenu des documents. Nous avons pour cela défini deux critères de pertinence que nous considérons durant cette étape d'enrichissent afin de sélectionner pour chaque document les jetons d'enrichissement pertinents. Nous étudions dans la section prochaine l'impact de ces critères sur la qualité des résultats de recherche.

B. Évaluation de l'impact du contexte de recherche et de la représentativité des données d'intérêt lors de la description des documents, sur le processus de RI

Afin de valider l'importance des deux critères de pertinence sur lesquels se base la description personnalisée des documents, le protocole d'évaluation suivant est appliqué :

- Nous appliquons deux techniques de description des documents :
  - La première technique est la description personnalisée proposée par notre modèle. Elle prend en considération les deux critères de pertinence d'un jeton d'enrichissement durant le processus de description personnalisée des documents, à savoir son contexte et sa représentativité pour le contenu de chaque document. Cette description exploite les requêtes et les étiquettes des utilisateurs pour enrichir l'index multidimensionnel des documents (cf. section V.2.1).
  - La deuxième technique met en œuvre le modèle social BM25FS (Bouhini 2014) qui exploite uniquement les étiquettes des utilisateurs pour enrichir le contenu des documents au sein d'un seul espace d'indexation sans tenir en compte des deux critères de pertinence qui sont définis par notre modèle.
- Nous évaluons ensuite le processus de RI à travers des requêtes ambiguës. L'ambigüité d'une requête de recherche est validée lorsque plusieurs sujets de recherche sont attribués à cette requête dans les profils des utilisateurs. Il consiste à évaluer l'efficacité du système à répondre pertinemment aux besoins des utilisateurs dans les deux cas précités relatifs aux deux techniques de description des documents. Les résultats de cette évaluation sont présentés à travers la figure 7.29.

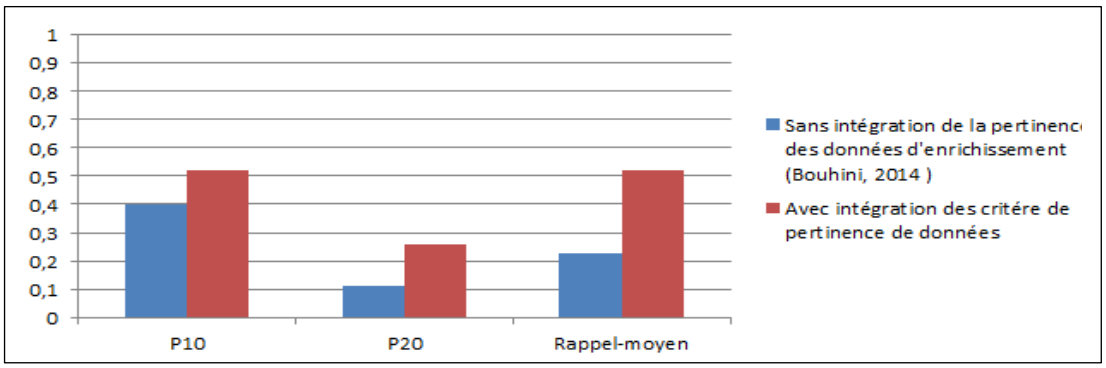


Figure 7. 29. Évaluation de l'importance du contexte et de la représentativité des données dans l'enrichissement et la recherche de documents

Les résultats des précisions P10, P20 et MAP du système qui sont illustrés dans la figure 7.29 confirment les observations théoriques présentées dans le chapitre 5 (cf. section V.2.2.5). En effet, lorsque le contexte et la représentativité des données d'enrichissement ne sont pas prises en compte, ces données peuvent être non pertinentes pour le contenu de certains documents et aura un effet négatif sur le processus de RI. Ceci explique une faible précision avec le modèle de Bouhini (Bouhini 2014).

Notre modèle de référence obtient également de faibles résultats en rappel. Ceci s'explique par le fait que ce modèle se base sur une correspondance syntaxique entre le contenu de la requête et celui des documents dans l'index. De cette façon, même lorsque les intérêts des utilisateurs sont intégrés dans la description de ces documents, cela est insuffisant pour localiser les documents qui sont sémantiquement liés aux intérêts de l'utilisateur. Ceci est comblé dans notre modèle par la considération d'un univers de description multidimensionnel qui prend en considération la sémantique des documents. La prochaine section aborde, avec plus de détail, l'avantage de cet univers de description multidimensionnel.

C. Évaluation de l'impact d'un univers de description multidimensionnel des documents sur leur enrichissement et sur la pertinence de la RIP.

Nous avons soulevé dans la section V.2.2.5 (page 190), l'avantage d'un univers de description multidimensionnel des documents sur le processus d'enrichissement orienté-utilisateur de leur contenu. Cet avantage se traduit par une diversité d'enrichissement qui permet d'enrichir également les documents qui sont sémantiquement liés aux intérêts de l'utilisateur. Un exemple illustratif a été présenté dans la figure 5.8 du chapitre 5.

Afin de démontrer l'utilité de cet univers multidimensionnel dans l'amélioration de la RI, les résultats obtenus de notre modèle multidimensionnel sont comparés au modèle classique où i) les documents sont décrits uniquement avec leur contenu textuel dans un seul espace monodimensionnel et ii) l'enrichissement personnalisé des documents se base sur une simple comparaison syntaxique entre les intérêts de l'utilisateur et les jetons constituant le contenu de ces documents.

La comparaison est effectuée en fonction du rappel système. Cette métrique est choisie, car le but est d'évaluer la capacité du système à détecter tous les documents pertinents qui existent lorsque la sémantique des jetons d'indexation est considérée dans la description et l'enrichissement des documents. Dans cette étape d'évaluation, nous suivons le protocole suivant :

- Préparation des requêtes de test associées aux utilisateurs. Pour ce faire, les intérêts des utilisateurs sont organisés en sujet d'intérêt et une requête est sélectionnée de chaque sujet, cela permet d'évaluer le système à différents besoins. Le reste des données d'intérêt est utilisé pour la description des documents.
- Ensuite, nous augmentons explicitement le contenu des profils utilisateurs avec des documents qui sont sémantiquement liés au contenu des requêtes de test préparées. Puis nous testons si les deux modèles que nous mettons en comparaison sont capables de les sélectionner. Pour mieux expliquer cette étape, prenons l'exemple de la figure 5.8. Ainsi, pour un utilisateur qui s'est intéressé à « sofa » et « divan » lors de ses recherches antérieures, des documents annotés avec canapé et/ou causeuse et/ou fauteuil sont ajoutés dans son profil. Puis, les requêtes « ventes de sofa » ou « achat de divan » ou simplement «sofa» ou « divan » sont exploitées pour interroger le système et tester si les documents qui sont ajoutés explicitement dans les profils des utilisateurs peuvent être sélectionnés.

	Avec description multidimensionnelle	**Avec description monodimensionnelle**
Rappel-moyen	0.57	0.21

Tableau 7. 8. Pourcentage d'amélioration du rappel système lorsqu'un univers de description multidimensionnel est adopté au lieu d'un univers monodimensionnel.

Nous pouvons voir à travers les résultats du tableau 7.8, que le modèle avec une description multidimensionnelle des documents donne un rappel meilleur au système que lorsqu'une description monodimensionnelle est adoptée. Cette dernière se base sur une description personnalisée des documents

à base de données d'intérêt des utilisateurs qui correspondent syntaxiquement à leur contenu textuel. Cet enrichissement est insuffisant et ne permet pas de localiser les documents qui ne sont pas décrits exactement avec la requête de l'utilisateur. Cela est comblé dans notre modèle. Si nous revenons à l'exemple de la figure 5.8 du chapitre 5, nous pouvons voir que les documents qui sont décrits dans l'espace identitaire avec « canapé » ou « fauteuil » sont également décrits dans l'espace sémantique avec « sofa ». Ainsi, lorsque le jeton «sofa» représente un besoin récurrent dans le profil de l'utilisateur, il permet de promouvoir aussi les documents précités, car l'espace sémantique a permis d'enrichir leur univers de description. Ce qui n'est pas le cas avec les modèles de Bouhini (Bouhini 2014) et de Bouadjenek (Bouadjenek *et al.* 2016).

Il a été prouvé dans plusieurs travaux de recherche que l'intégration du voisinage utilisateur permet d'améliorer la pertinence des résultats de recherche des utilisateurs (Bouhini 2014) (Bouadjenek *et al.* 2016). La section suivante traite cette direction de recherche.

D. Évaluation de l'impact du voisinage utilisateur sur le processus de RI

Le but de cette évaluation n'est pas d'évaluer l'importance du voisinage utilisateur dans le processus de RI, cette efficacité a été déjà prouvée au cours de plusieurs études de recherche (Bouhini 2014; Bouhini *et al.* 2016), mais l'objectif de nos expérimentations est de tester la technique que nous proposons pour créer les groupes d'intérêt des utilisateurs (cf. section V.2.1.2.4). Nous avons montré dans le modèle théorique comment les documents sont structurés en champs de description et comment le voisinage de l'utilisateur est pris en considération durant la recherche d'information. Il consiste à calculer le voisinage de l'utilisateur cible puis effectuer une recherche personnalisée qui cible les k-champs décrivant les documents selon l'utilisateur et son voisinage. Le protocole d'évaluation est comme suit :

- Calculer le voisinage de chaque utilisateur. Ce processus se base sur le calcul du groupe SFC qui couvre le contexte de la requête de recherche cible de chaque utilisateur.
- Sélectionner les requêtes de test de chaque utilisateur. Deux catégories de requêtes sont considérées :
    - La première catégorie est l'ensemble des requêtes qui représentent à la fois un besoin non fréquent chez l'utilisateur cible (faible fréquence) et un besoin récurrent chez un ou plusieurs utilisateurs de son voisinage. Cela permet de tester la capacité de ce voisinage à

augmenter le score de correspondance des documents d'intérêt de l'utilisateur lorsque ses voisins sont pris en considération durant la recherche.

- Un autre type de requête est exploité, il s'agit des requêtes qui représentent à la fois un nouveau besoin chez l'utilisateur et un besoin fréquents chez son voisinage.

- Les résultats de recherche obtenus de cette intégration du voisinage utilisateur sont comparés aux résultats où seuls les intérêts individuels des utilisateurs sont considérés dans la recherche. Cette comparaison est présentée en termes d'amélioration en rappel moyen et en précision moyenne du système.

	Avec intégration du voisinage dans la recherche d'information
Amélioration en rappel-moyen	11%
Amélioration en précision moyenne (MAP)	23%

Tableau 7. 9. Résultats de la pertinence des résultats de recherche du système lorsque le voisinage de l'utilisateur est intégré

L'intégration du voisinage utilisateur a donné des améliorations au système de 17% en rappel et de 23% en précision. Ces améliorations s'expliquent comme suit :

- L'amélioration en précision est justifiée par le fait que lorsque les documents sont enrichis uniquement avec les intérêts individuels des utilisateurs, ceux qui sont liés à un besoin non récurrent de l'utilisateur cible ne sont pas fortement influencés durant ses recherches. Cependant, lorsque la recherche de l'utilisateur représente un besoin récurrent chez un ou plusieurs utilisateurs de son voisinage, l'implication de ces utilisateurs voisins dans la description de ces documents permet d'influencer ces documents en augmentant le nombre d'occurrences des jetons dans leur description. Ceci augmente à son tour le score de correspondance de ces documents avec la requête lors de la recherche, ce qui leur permet de gagner en classement et augmente ainsi la précision P@X du système (le nombre de documents pertinents sur les X premiers rangs).
- Lorsque la recherche de l'utilisateur cible correspond à un nouveau besoin en information pour lui et représente en même temps un besoin fréquent chez un ou plusieurs utilisateurs voisins, l'implication des intérêts de ce voisinage dans la description des documents permet de sélectionner les documents qui ne peuvent pas être localisés lorsqu'ils sont décrits uniquement à travers les intérêts de l'utilisateur cible. Ceci explique l'amélioration dans le rappel du système.

Le voisinage de l'utilisateur évolue au fur et à mesure que les intérêts de l'utilisateur évoluent. Un des facteurs qui influence négativement sur la pertinence des résultats du système est la non-considération de l'évolution des intérêts de l'utilisateur. Ainsi, lorsque le voisinage de l'utilisateur n'est pas pertinent, c'est-à-dire, il ne représente pas les besoins en information courants de l'utilisateur, cela ne permet pas de donner une amélioration au système. Il est donc important d'évaluer la pertinence des résultats lorsque le groupe d'intérêt de l'utilisateur évolue au fil du temps mais n'est pas pris en considération par le système. Cette évaluation est discutée dans la prochaine section.

E. Évaluation de l'importance de la dynamicité du voisinage utilisateur

Cette section présente l'évaluation de l'importance de la dynamicité du voisinage utilisateur dans la description des documents et lors du processus de RI. Le protocole d'évaluation de cette tâche est effectué comme suit :

Nous calculons en premier lieu les groupes d'intérêt des utilisateurs à différentes périodes de temps et sélectionnons en suite les utilisateurs à qui le système a pu attribuer différents groupes d'intérêt (voisinages) dans les différentes périodes considérées. Nous jugeons que cet ensemble d'utilisateurs est bien adapté pour valider l'importance de cette dynamicité du voisinage utilisateur. Pour ce faire, nous adoptons deux techniques d'enrichissement personnalisé des documents :

- La première considère le voisinage de chacun de ces utilisateurs sélectionnés comme une entité statique qui ne change pas au fil du temps en l'exploitant à la fois dans la description des documents, et tout au long des recherches de ces utilisateurs (Bouhini 2014).
- La deuxième considère la dynamité du voisinage utilisateur en adoptons une description documentaire personnalisée à base de champs de description. Cette technique utilise un champ de description pour décrire chaque document selon chaque utilisateur. Cela permet d'effectuer une recherche personnalisée sur les k-champs relatifs au voisinage de chaque utilisateur. Ce voisinage est dynamique et est calculé à chaque période de temps.

Nous sélectionnons ensuite pour chacun de ces utilisateurs un ensemble de requêtes de test qui sont exploitées pour évaluer les deux techniques précitées. Le choix de ces requêtes est basé sur le même protocole présenté dans la section précédente, c'est-à-dire, une requête est sélectionnée de chaque sujet

d'intérêt. Les résultats de recherche obtenus de ces deux techniques d'enrichissement sont présentés en termes de pourcentage d'amélioration du système de recherche par rapport au cas où seuls les intérêts des utilisateurs sont prise en considération.

	Sans dynamicité du voisinage	Avec dynamicité du voisinage
Amélioration en rappel-moyen	0%	13%
Amélioration en précision moyenne (MAP)	0%	19%

Tableau 7. 10. Évaluation de l'impact de la dynamicité du voisinage sur les recherches de l'utilisateur cible

Les résultats du tableau 7.10 montrent que lorsque le système ne prend pas en considération l'évolution du voisinage, ce dernier n'aura pas d'impact sur les recherches de l'utilisateur. On voit qu'aucune amélioration ni en rappel (0%) ni en précision (0%) n'est apportée lorsque la dynamicité du voisinage est négligée. Tandis qu'avec la prise en compte de cette dynamicité, le système de recherche est amélioré de 13% et 19% respectivement en rappel et en précision.

F. Évaluation de l'impact d'une structure à base de champs de description sur l'efficience du système

Cette section présente l'évaluation de la technique adoptée par notre système pour personnaliser la description des documents. Cette technique de personnalisation se base sur les champs de description où chaque document est décrit avec plusieurs champs qui décrivent son contenu selon chaque utilisateur. Cette structure permet d'effectuer des recherches personnalisées qui ciblent un sous-ensemble de champs de contenu qui décrivent les documents selon l'utilisateur cible et son voisinage. Ceci vise à optimiser l'espace mémoire exploité par le système pour la description personnalisée des documents (cf. section V.2.2.5 page 189). Nous évaluons ainsi la taille de l'univers d'indexation de notre système selon l'évolution du nombre d'utilisateurs et le nombre de documents, puis le comparons avec la proposition de Bouhini et ses collègues qui représente le système de référence de cette tâche d'évaluation (Bouhini 2014). Ces auteurs proposent de créer pour chaque utilisateur un index documentaire à part qui décrit les documents selon les un utilisateur et son voisinage.

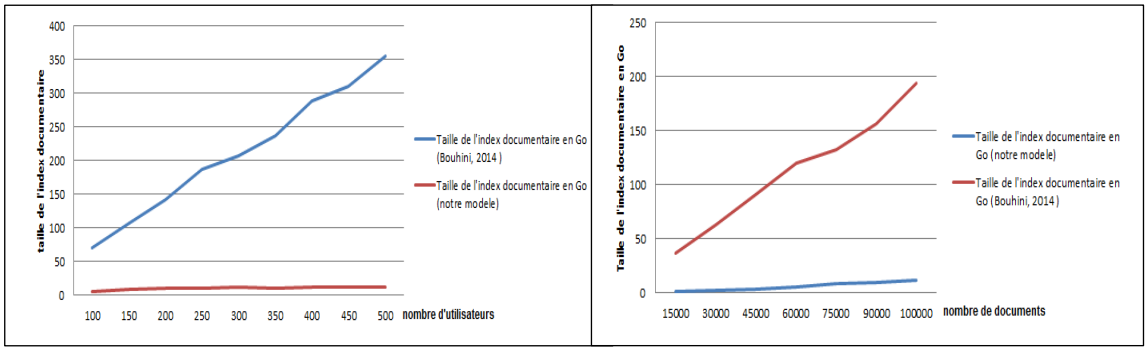


Figure 7. 30. Impact du nombre d'utilisateurs et du nombre de documents sur la taille de l'index documentaire personnalisé de notre système et celui du système de référence (Bouhini 2014)

Les résultats de la figure 7.30 montrent que l'évolution des documents et des utilisateurs système a un impact plus marqué sur la taille de l'index documentaire lorsque la technique de description des documents proposée par le système de référence (Bouhini 2014), est adoptée. Plus concrètement, avec 100 utilisateurs et 100 milles documents, la taille de l'index documentaire obtenue avec le système de référence est de 192 giga octets. Elle est beaucoup plus supérieure que la taille obtenue avec notre technique de description qui est de 11.7 giga octets. En outre, on remarque que lorsque le pas d'incrémentation du nombre des utilisateurs est de 500 utilisateurs, l'augmentation de la taille de l'index s'élève de 120 % dans notre système et de 401% dans le système de référence. Aussi, lorsque le pas d'incrémentation des documents est de 15000 documents, la taille de l'index augmente de 45 % dans notre système et de 94.3 % dans le système de référence. Ces résultats valident les anticipations théoriques présentées dans le modèle théorique du chapitre 5 (cf. section V.2.2.5 page 190).

VII.5.3.2 Évaluation du modèle de personnalisation par intégration du profil utilisateur au niveau du réordonnancement des résultats

Ce modèle personnalise les recherches de l'utilisateur en exploitant ses intérêts et ceux de son voisinage dans le réordonnancement des documents résultants. Le principe de ce modèle consiste à identifier le besoin informationnel de l'utilisateur derrière sa requête de recherche et le projeter sur le contenu de son profil et ceux des autres utilisateurs voisins en vue d'extraire les centres d'intérêt qui couvrent ce besoin en information. Ces centres d'intérêt sont exploités pour la personnalisation des résultats. Lorsque la requête de l'utilisateur est ambiguë, l'exploitation de son profil aide à désambiguïser son contenu en identifiant parmi les interprétations possibles celle qui représente un intérêt pour cet utilisateur. Cette

solution n'est pas nouvelle, elle a été adoptée au cours de plusieurs études (Rose et Levinson 2004) (Liu *et al.* 2006; Luo *et al.* 2014). La solution apportée par notre modèle vise à résoudre un problème un peu plus complexe. Ce problème se traduit par l'existence de plusieurs interprétations (sujets d'intérêt) dans le profil de l'utilisateur auxquels est liée la requête de recherche. Une technique d'identification de l'intention utilisateur est alors proposée. Elle consiste à prédire parmi les différents centres d'intérêt identifiés dans son profil celui qui représente son besoin informationnel courant (cf. section V.2.2 page 192). Le centre d'intérêt identifié est exploité pour personnaliser les résultats de recherche. Cette section englobe deux sous taches d'évaluation qui sont liées à ce modèle, elles s'énumèrent comme suit :

- Évaluation de l'efficacité de la technique d'identification du besoin informationnel de l'utilisateur au sein de son profil lorsque sa requête de recherche est ambigüe.
- Évaluation de l'efficacité de la technique hybride de réordonnancement des résultats en la comparant à la fois au modèle de la RI classique proposé dans la section VII.4 et au modèle personnalisé proposé dans la section précédente (cf. section VII.5.3.1) qui intègre les intérêts de l'utilisateur dans la description des documents.

A. Évaluation de la technique d'identification du besoin utilisateur au sein de son profil

L'identification du besoin informationnel de l'utilisateur consiste à conceptualiser la requête utilisateur en projetant son contenu sur le contenu d'une ontologie de référence (Dmoz), puis identifier dans le profil de cet utilisateur le sujet d'intérêt qui corrèle avec la représentation conceptuelle obtenue de la requête. Afin de mettre en œuvre et évaluer cette technique, le protocole illustré dans la figure 7.31 est appliqué, il englobe les étapes suivantes :

Étape 1. **Conceptualisation de la requête utilisateur.** Il consiste à indexer au préalable le contenu de l'ontologie Dmoz où chaque document est décrit avec son contenu textuel et un concept auquel il est associé dans la prédite ontologie, puis exploiter l'index conçu pour des recherches full texte. Pour chaque requête de recherche, le résultat est un ensemble de n documents qui correspondent à son contenu, et la représentation conceptuelle de cette requête est le vecteur conceptuel constitué des k concepts qui correspondent aux documents résultants. Chaque concept est pondéré avec un score qui représente la moyenne des scores de correspondance de la requête utilisateur avec le sous-ensemble de documents

résultants qui sont associés à ce concept. Cette étape dépend du nombre de documents à prendre en considération dans la représentation conceptuelle de cette requête. Ce nombre est important pour l'extraction des concepts pertinents qui aident à identifier le centre d'intérêt qui couvre le besoin informationnel de l'utilisateur.

Étape 2. Définition du nombre optimal des documents. Cette étape se base sur une phase d'apprentissage où l'impact du nombre des documents est étudié sur l'identification des sujets d'intérêt qui couvrent la requête de l'utilisateur dans son profil. Pour ce faire, nous exploitons un ensemble de requêtes d'apprentissage extraites des profils des utilisateurs. Ces requêtes sont déjà associées à des sujets d'intérêt. Nous appliquons à chaque requête la première étape de conceptualisation puis évaluons au cours de la prochaine étape (étape 3) la capacité du système à identifier les sujets appropriées lorsque le nombre de documents varie de 5 à 60 (cf. figure 7.32). Cette évaluation se base sur le calcul de précision à chaque valeur de N, cette précision est définie dans la métrique suivante :

$$Précision = \frac{nombre\ de\ requetes\ correctement\ classées}{nombre\ de\ requetes\ d'apprentissage}$$

Étape 3. Extraction des centres d'intérêt qui correspond à la requête de l'utilisateur dans son profil. La requête utilisateur est corrélée à chaque sujet dans le profil de l'utilisateur. Cette corrélation se base sur l'exploitation de la mesure de Kendall qui mesure la correspondance entre le vecteur conceptuel pondéré de cette requête, obtenu dans la première étape, et le vecteur conceptuel pondéré de chaque centre d'intérêt dans le profil de l'utilisateur. Le résultat est le ou les centres d'intérêt qui répondent au seuil de corrélation définie dans la section VII.5.1.B.

Étape 4. Identification du besoin utilisateur. Cette étape consiste à identifier le besoin de l'utilisateur lorsque le contenu de sa requête est lié à plusieurs sujets d'intérêt dans son profil. Il consiste à sélectionner parmi les centres d'intérêt résultant de l'étape 3 celui qui couvre ses attentes. La technique de prédiction du besoin utilisateur est appliquée. Elle est évaluée en sélectionnant un ensemble de requêtes ambiguës depuis les profils des utilisateurs qui sont pré-annotées de sujets. Il s'agit de calculer la capacité du système à prédire les sujets d'intérêt adéquats aux besoins des utilisateurs exprimés par les requêtes sélectionnées. Cette capacité est évaluée à travers la métrique de précision définie ci-haut.

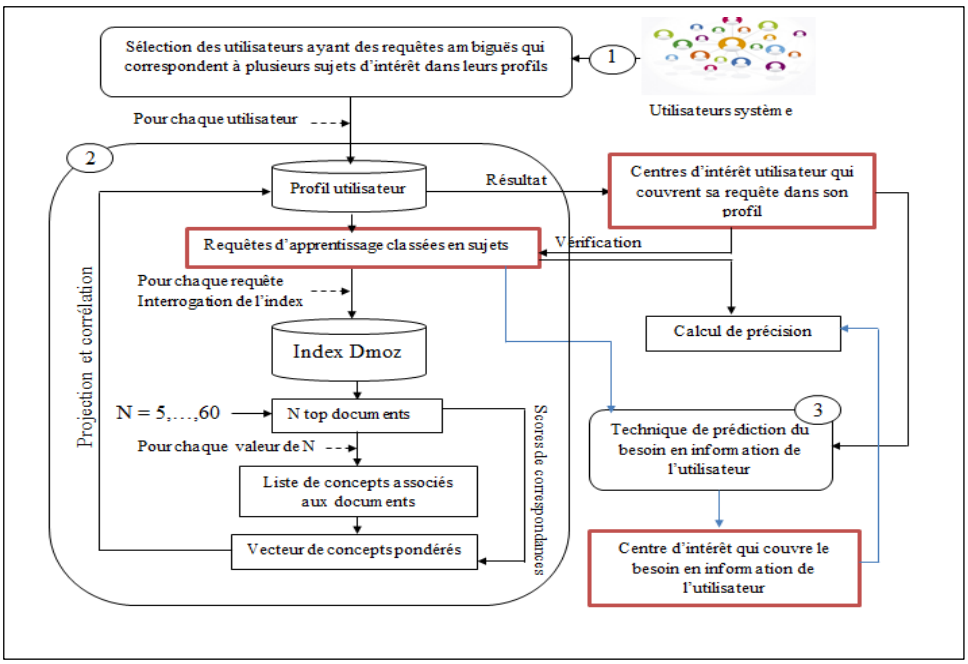


Figure 7. 31. Protocole d'évaluation du processus d'identification du besoin en information de l'utilisateur

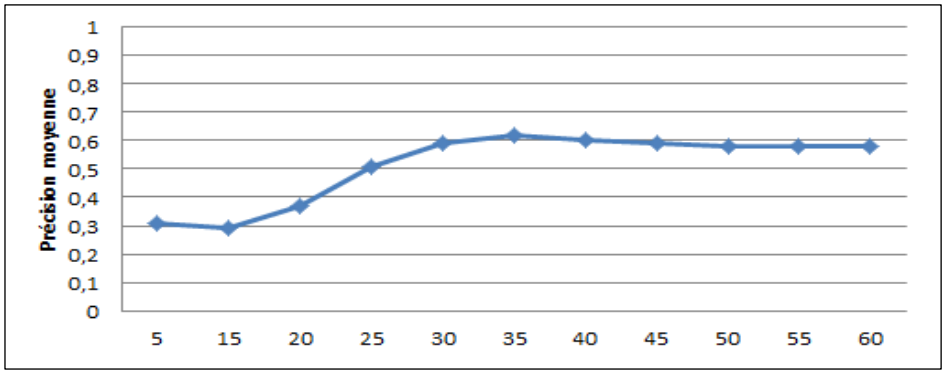


Figure 7. 32. Définition du nombre optimal de documents pour la détection du sujet d'intérêt utilisateur derrière une requête de recherche dans son profil

D'après les résultats de la figure 7.32, la meilleure précision est obtenue avec 35 documents. Ce nombre est utilisé pour le reste des expérimentations.

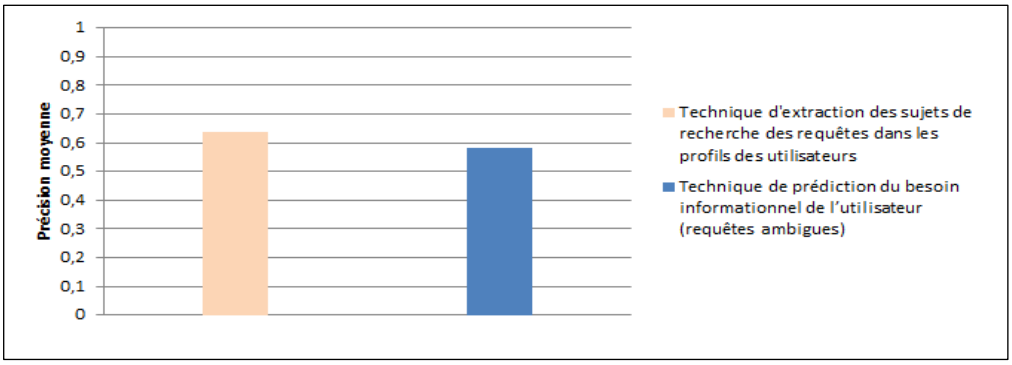


Figure 7. 33. Évaluation de la technique d'identification du besoin informationnel de l'utilisateur derrière une requête ambigüe

Les résultats obtenus dans la figure 7.33 sont promoteurs. Lorsque les requêtes de recherche des utilisateurs sont liées à des besoins en information récurrents, le système arrive à identifier dans leurs profils les sujets d'intérêt qui couvrent ces besoins. Cette technique a obtenu une précision de 0.61. Lorsque plusieurs sujets d'intérêt sont identifiés dans leurs profils derrière leurs requêtes de recherche, la technique de prédiction du besoin informationnel obtient une précision de 0.57.

B. Évaluation du modèle hybride de réordonnancement des résultats

Ce modèle exploite le centre d'intérêt de l'utilisateur qui couvre sa requête de recherche pour extraire les centres d'intérêt des autres utilisateurs voisins qui couvrent le même besoin en information, puis exploite le tous pour réordonner les documents résultants de cette recherche classique (cf. figure 7.33). Cette personnalisation se base sur une fonction linéaire qui combine trois scores de pertinence évaluant la pertinence globale d'un document (cf. équation 5.23), à savoir le score de correspondance classique du document et les scores de correspondances avec les intérêts de l'utilisateur et ceux de son voisinage. Cette pertinence globale dépend de trois coefficients de pondération α, β, ω qui définissent les degrés d'importance de chacun des facteurs pris en considération dans l'évaluation de cette pertinence globale, tel que $\alpha+ \beta+ \omega =1$ où :

- α est le degré d'importance de la correspondance thématique du document avec la requête dans le calcul de la pertinence globale du document,
- β est le degré d'importance de la correspondance personnalisée du document avec les données d'intérêt de l'utilisateur cible dans le calcul de cette pertinence. Il est égal au score d'intérêt de

l'utilisateur pour le sujet de recherche qui couvre la requête cible. Ce degré est stocké dans son profil et accumulé durant ses activités de recherche

- ω est le degré d'importance de la correspondance personnalisée du document avec les données d'intérêt du voisinage utilisateur. Il est égal au score d'intérêt de l'utilisateur pour le groupe d'intérêt SFC_i qui reflète son groupe s'intérêt courant.

Les résultats obtenus sont comparés aux modèles de références citées ci-haut. Pour cette évaluation, un protocole d'évaluation est proposé, il est mis en œuvre en 7 étapes qui sont résumées dans la figure 7.34.

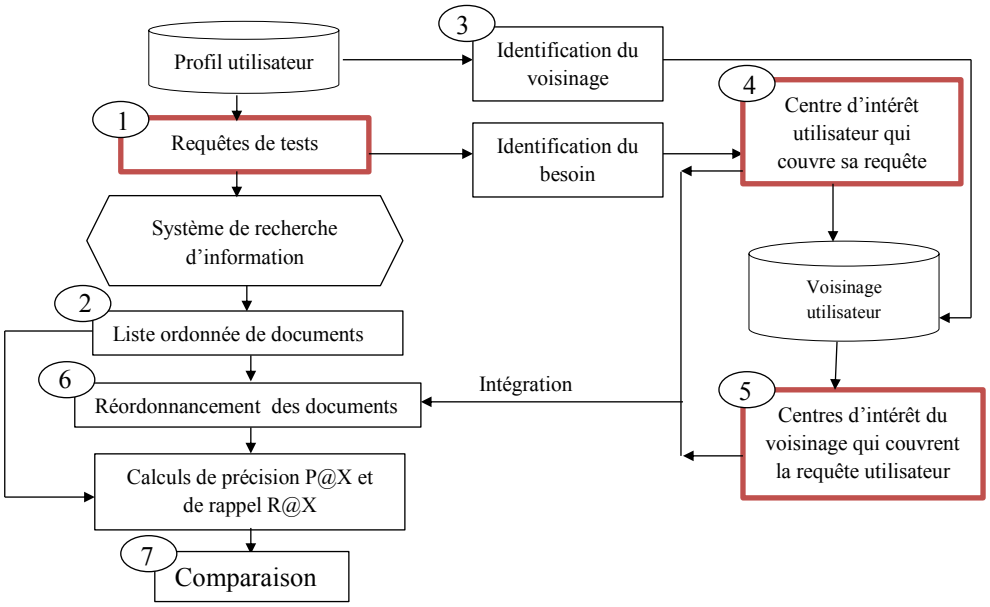


Figure 7. 34. Protocole d'évaluation du processus de personnalisation à base de réordonnancement contextuel des résultats de recherche

Les résultats obtenus de cette évaluation sont présentés dans la figure ci-dessous (cf. figure 7.35). Ils représentent une comparaison entre le système de réordonnancement contextuel des documents avec i) le système classique qui ne tient pas compte de l'utilisateur durant la recherche et avec ii) le système personnalisé qui intègre les données d'intérêts dans la description personnalisée des documents.

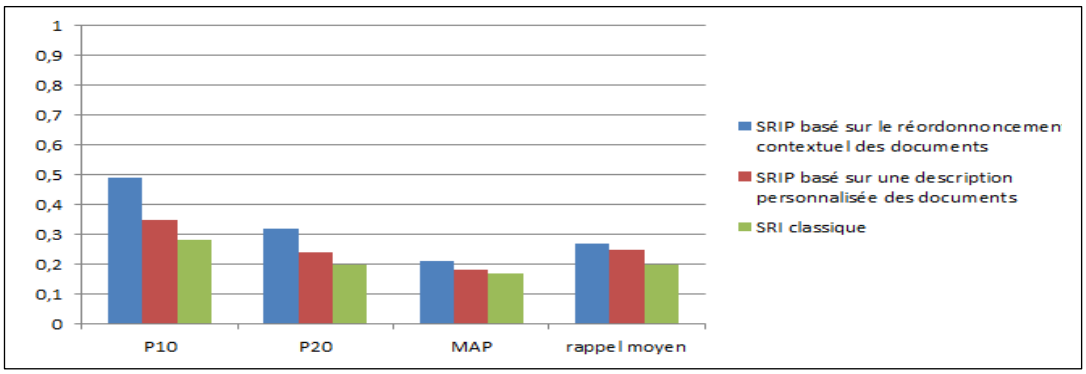


Figure 7. 35. Évaluation du modèle hybride de réordonnancement contextuel des résultats de recherche

Les résultats montrent que lorsque les données d'intérêt des utilisateurs sont prises en considération dans le réordonnancement des documents, le système de RI est meilleur en précision et en rappel. Cela peut être justifié comme suit :

- Par rapport à la recherche classique, le réordonnancement contextuel des documents permet de réajuster les scores de pondération des documents résultants d'une recherche classique avec leur degré obtenu par correspondance avec le contexte de la requête cible. Ce contexte est représenté par un ensemble de documents qui couvrent le sujet de recherche de cette requête dans le profil de l'utilisateur cible et les profils de ses utilisateurs voisins. Cela permet d'augmenter le score de correspondre des documents ayant un faible score de correspondance suite à une recherche classique, et leur permet de gagner en classement et améliore alors la précision P@X,
- Nous avons vu dans la section VII.5.3.1 que lorsque les données d'intérêt des utilisateurs (notamment leurs requêtes de recherche et leurs étiquettes d'annotation) sont intégrées dans la description des documents, cela permet d'augmenter le score de correspondance de ces documents avec les requêtes de recherche, en particulier lorsqu'ils couvrent syntaxiquement ou sémantiquement les intérêts des utilisateurs. Lorsque le contexte de la requête est exploité dans le réordonnancement des documents, cela permet de réajuster le score de pertinence des documents qui correspondent au sujet de recherche de la requête. Cette correspondance est plus générale que les deux correspondances, syntaxique et sémantique, prise en compte dans la première approche. Cela explique une meilleure pertinence de résultats notamment en précision P10 et P20 du SRIP (cf. figure 7.35).

- Un autre facteur qui peut être responsable de cette amélioration est le type de données d'intérêt qui sont prises en considération dans le processus de personnalisation des résultats. Par rapport aux requêtes et étiquettes qui sont exploitées dans la description personnalisée des documents, les documents qui sont exploités dans le réordonnancement des résultats représentent un contenu plus riche qui aide à mieux personnaliser les résultats de l'utilisateur en détectant la similarité des documents résultants de la recherche avec les documents d'intérêt de l'utilisateur, ceci aide à améliorer la précision dans les N premiers documents.

C. Évaluation de l'approche de proposition de nouveaux documents à base de sujet de recherche et de voisinage utilisateur

Cette approche propose à l'utilisateur en réponse à sa recherche, les documents qui n'ont pas été déjà consultés dans ses sessions antérieures en se basant sur les expériences des autres utilisateurs voisins, en particulier lorsque sa requête est liée à des recherches antérieures. Ces documents couvrent le même sujet de la recherche courante et sont extraits des profils des autres utilisateurs voisins (cf. section V.2.2.1). Cette recherche contextuelle par sujet de recherche est plus générale qu'une recherche par mots clés, et peut être utile pour améliorer les résultats de l'utilisateur. Afin d'évaluer l'efficacité de cette approche, un protocole d'évaluation est suivi et se résume comme suit :

- Calculer le voisinage de chaque utilisateur. Ce processus se base sur le calcul du groupe SFC pertinent en se basant sur le contexte de la requête, puis le calcul des utilisateurs voisins depuis le cluster des utilisateurs ayant le même intérêt pour ce groupe SFC (cf. figure 5.11).
- Sélectionner les utilisateurs ayant des documents en commun avec leurs utilisateurs voisins, ces utilisateurs représentent les utilisateurs de test.
- Supprimer depuis les profils de ces utilisateurs de test, les documents qui représentent des éléments en commun avec leurs voisinages.
- Tester la capacité du système à sélectionner les documents supprimés.
- Renvoyer les documents en liste ordonnée par leur score de pertinence qui représente leur popularité chez l'utilisateur cible estimée par le système (cf. équation 5.25). L'évaluation est donc basée sur le calcul de précisions P10, P20, MAP et de rappel moyen.

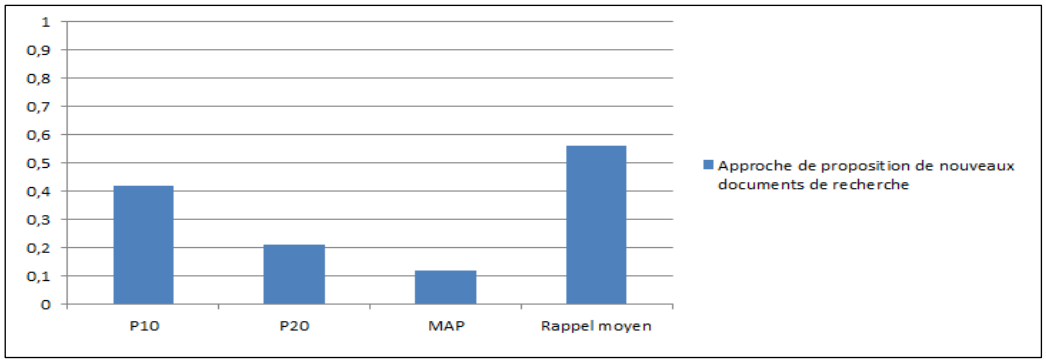


Figure 7. 36. Évaluation de l'approche de proposition de nouveaux documents de recherche

Les résultats de la figure 7.36 montrent que le système arrive à sélectionner les documents supprimés et obtient un bon rappel. Par ailleurs, les valeurs de précisions obtenues montrent que le système offre une valeur satisfaisante en précision P10. Cette précision diminue en P20 et est faible en MAP. Nous jugeons que la précision P10 est suffisante, car l'utilisateur a tendance à prendre plus d'attention aux 10 premiers documents.

VII.6. Exemple récapitulatif

Pour récapituler les principales contributions de cette thèse, prenons l'exemple d'une collection de documents appartenant à différents domaines de recherche. Ces documents sont représentés par notre système à travers trois espaces de représentation (cf. figure 7.37).

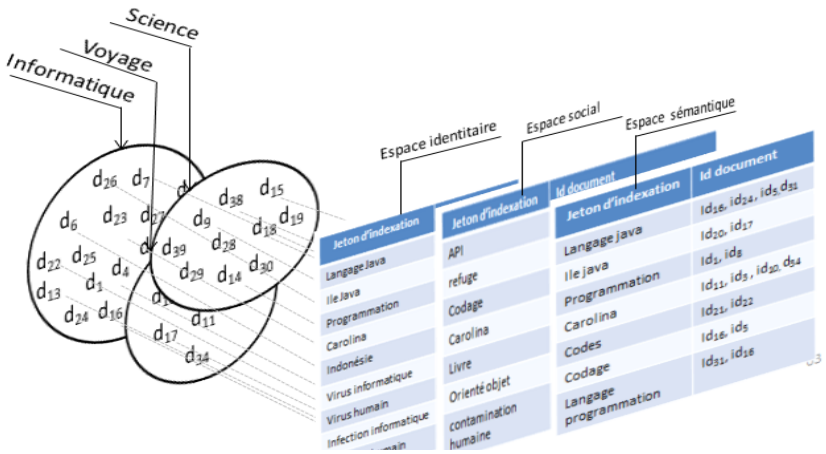


Figure 7. 37. Représentation multidimensionnelle d'une collection de documents web

Lorsqu'une requête de recherche est envoyée au système, telle que la requête « tutoriel programmation Java », le système i) interprète la requête en extrayant les jetons représentatifs, ii) traite l'ambiguïté des jetons polémiques, puis iii) prépare les dimensions d'enrichissement de cette requête. Les différentes interprétations de la requête dans l'index sont renvoyées sur différentes facettes de données :

- La première facette de données (la facette identitaire) répond aux besoins des utilisateurs qui ont une préférence pour un contenu qui correspond exactement leurs requêtes de recherche, notamment lorsqu'ils sont à la recherche d'un titre d'un livre, d'un article scientifique, etc.
- La deuxième facette de données (la facette sociale) répond aux besoins des habitués des réseaux sociaux. Les résultats sont beaucoup plus structurées et informatifs à travers les étiquettes des utilisateurs.
- La troisième facette (la facette sémantique) répond aux utilisateurs qui ont besoin de naviguer d'autres résultats connexes sur d'autres langages de programmation.

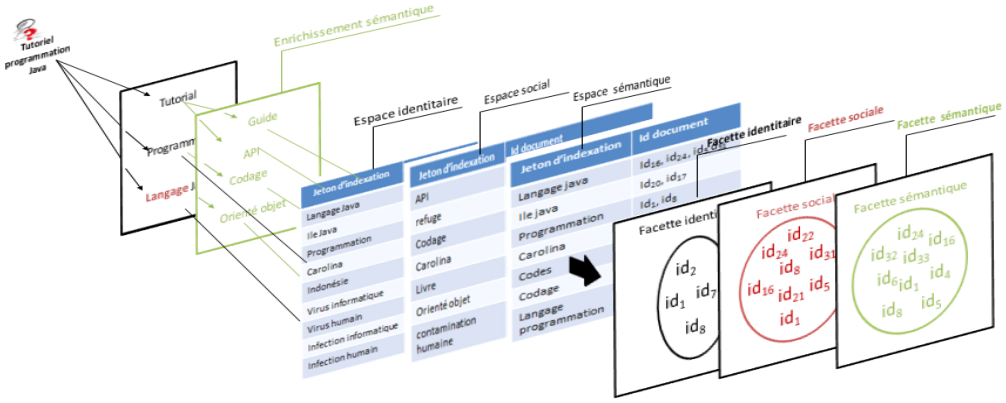


Figure 7. 38. Recherche multidimensionnelle de documents web

Avec un système monodimensionnel les documents sont représentés par une seule vue de données (cf. figure 7.39). Lorsque la requête est envoyée au système, les résultats de recherche sont renvoyés mélangés sur une seule vue d'affichage sans laisser la possibilité à l'utilisateur de distinguer entre les différentes interprétations possibles.

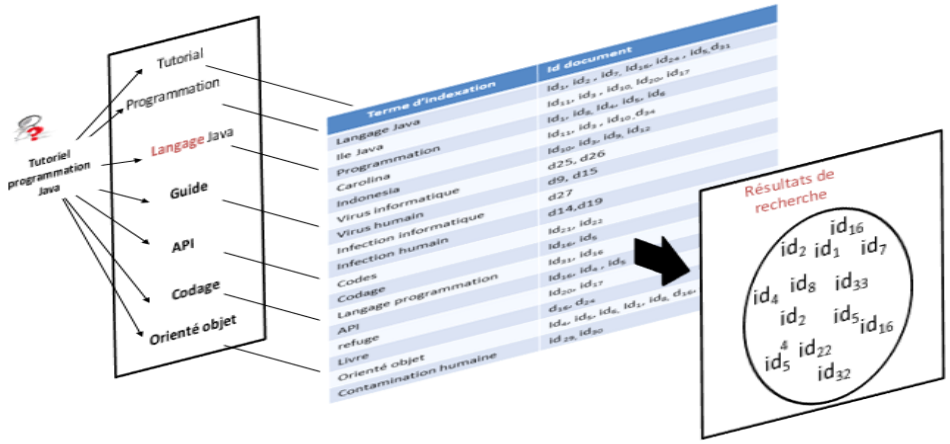


Figure 7. 39. Représentation et recherche avec un système monodimensionnel

Prenons maintenant l'exemple d'une requête ambiguë, telle que la requête « types de virus » (cf. figure 7.40), comme déjà abordé dans un précédant exemple cette requête est liées à plusieurs interprétations, notamment les virus humain et les virus informatique. Pour répondre à cette requête, le système s'appuie sur le profil de l'utilisateur pour proposer à l'utilisateur une vue personnalisée qui répond à des besoins plus spécifiques. Alors, pour un l'utilisateur qui a déjà effectué des recherches liées aux virus informatique ou à des sujet connexes, telle que la cryptologie ou les politiques de sécurité (le cas de l'utilisateur 1 dans la figure 7.40), cette vue va promouvoir les documents qui couvrent les virus informatiques. Sinon si la requête représente un nouveau besoin pour l'utilisateur (le cas de l'utilisateur 2), le système se base sur son voisinage pour lever l'ambigüité sur cette requête.

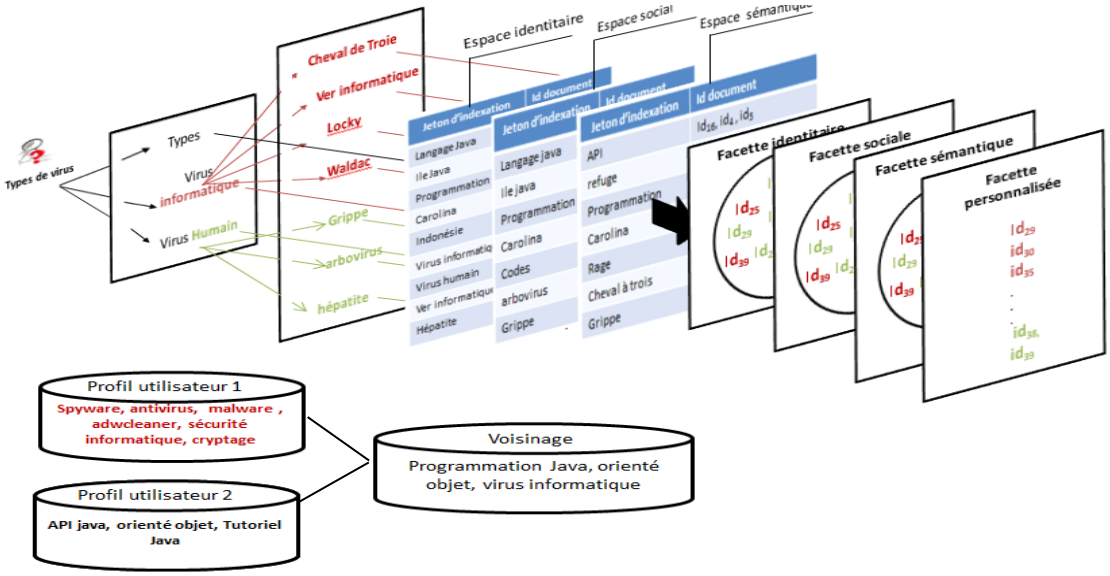


Figure 7. 40. Recherche d'information multidimensionnelle avec une requête ambiguë

VII.7. Conclusion

Nous avons présenté dans ce chapitre la mise en œuvre de notre système de recherche d'information qui se base sur une nouvelle technique d'indexation et de recherche d'information multidimensionnelle. Cette nouvelle technique se base sur l'exploitation de différents aspects de représentation du contenu, ils sont rendus transparents pour l'utilisateur dans le but d'améliorer la visibilité de ses résultats. La mise en œuvre de cette technique a permis de valider l'efficacité des facettes de données et des valeurs de facettes proposées par le système. Les expérimentations ont montré que l'interface multidimensionnelle proposée permet de mieux répondre aux différents besoins en information des différents utilisateurs et est plus efficace que les deux recherches, traditionnelle et monodimensionnelle. Cette satisfaction a permis de valider à son tour l'efficacité des différents aspects (syntaxique, sémantique et social) qui sont exploités et combinés ensemble au sein de notre système pour la représentation des documents et l'interprétation des requêtes utilisateur. Cette hybridation a contribué à l'amélioration de la recherche utilisateur.

Par ailleurs, lorsque la requête de l'utilisateur englobe un jeton polysémique, la technique de désambiguïsation proposée a permis d'améliorer la précision du système en écartant les documents qui ne couvrent pas le sujet de recherche auquel est liée la requête. Ces documents peuvent être retournés lorsqu'un seul jeton de la requête correspond à leur contenu, en particulier le jeton polysémique.

Ce chapitre a permis de valider également l'efficacité du profil utilisateur, et plus particulièrement sa représentation multidimensionnelle, dans l'amélioration de la recherche utilisateur. Cette amélioration a été discernée au cours de plusieurs processus :

- Dans l'amélioration du processus de prédiction des intérêts des utilisateurs qui exploite différents niveaux de représentation de ces intérêts pour améliorer la corrélation entre les utilisateurs.
- Dans l'amélioration de la recherche des utilisateurs lorsque :
    - L'ambiguïté des données d'intérêt spécifiques des utilisateurs est levée à travers les sujets de recherche qui sont construits automatiquement en délimitant les activités de recherche par contexte de recherche.
    - Lorsque les données d'intérêt de ces utilisateurs sont prises en compte durant la recherche. Les expérimentations ont montré que lorsque ces données sont intégrées dans le réordonnancement des résultats cela a donné de meilleurs résultats par rapport à leur intégration au niveau de

l'indexation des documents. Dans cette étape d'évaluation, l'importance de deux critères de pertinence a pu être validée lors de la sélection des données d'enrichissement pour la description personnalisée des documents, notamment le contexte des données représenté par un sujet de recherche, et leur représentativité pour le contenu de ces documents. Ces critères ont aidé à améliorer la description et la découverte de ces documents.

- Lorsque des communautés d'intérêts sont construites à base de sujets fortement connexes, ceci a amélioré la recherche collaborative des utilisateurs.

Ces expérimentations ont permis également de valider l'efficacité des trois aspects contextuel, fréquentiel et temporal dans la gestion du contenu du profil utilisateur. L'intégration de ces différents facteurs a permis une exploitation pertinente des données stockées dans ce profil pour une meilleure pertinence de résultats de recherche.

Nous avons au cours de ce chapitre pu aussi valider l'efficacité de l'analyse comportementale hybride de l'utilisateur dans l'extraction de ses intérêts. Cette hybridation a eu un double avantage :

- Améliorer la détection des documents d'intérêt de l'utilisateur. Ces documents représentent l'élément fondamental pour la construction et l'enrichissement de son profil d'intérêts multidimensionnels.
- Améliorer les corrélations entre les utilisateurs en exploitant à la fois ses requêtes et ses étiquettes dans l'inférence collaborative des intérêts.

A travers ces expérimentations, la recherche à base de sujets de recherche a prouvé son efficacité dans l'amélioration de la recherche des documents. L'exploitation de ces sujets de recherche a permis d'améliorer :

- La sélection des règles d'inférence pertinentes pour les utilisateurs.
- La sélection personnalisée de ces règles d'inférence pour chaque utilisateur à travers les sujets fortement connexes.
- Le réordonnancement contextuel de résultats de recherche.
- L'identification du besoin en information de l'utilisateur lorsque sa requête est ambiguë et plusieurs centres d'intérêt sont identifiés dans son profil. Le système exploite les groupes des sujets fortement connexes pour sélectionner les données pertinentes depuis son profil.

- Lorsque le voisinage de l'utilisateur est défini à base des groupes de sujets fortement connexes. Cela a aidé à améliorer la recherche utilisateur lorsque les intérêts de ce voisinage sont pris en compte dans la description personnalisée des documents.

Cette thèse ouvre la voie à plusieurs perspectives. Ils sont discutés dans le chapitre suivant de la conclusion générale.

Chapitre 8 : Conclusion générale

Dans la présente thèse, nous avons présenté un nouveau système de recherche d'information qui fait appel à différents concepts de différents domaines en vue d'améliorer la recherche de l'utilisateur. Nous résumons ci-après les contributions, les limitations qui leurs sont liées ainsi que les futurs travaux.

VIII.1. Contributions

Les contributions de ce travail de recherche ont porté sur cinq volets :

1. Proposition d'un nouveau modèle d'indexation et de recherche d'information multidimensionnelle. Ce modèle a été proposé pour faire face à la surcharge d'information sur le web en améliorant la pertinence et la visibilité des résultats de recherche. Il se base sur une nouvelle technique d'indexation et de recherche d'information qui exploite différentes sources d'enrichissement sémantique et social. Ces sources sont distinguables par l'utilisateur en sein de son interface de recherche sous la forme de facettes de données et de valeurs de facettes afin d'améliorer les résultats des recherches. Cette distinction est rendue possible grâce au nouveau concept de projection multi-espaces proposé. Les facettes et leurs valeurs de données ont prouvé leur efficacité. Ils ont permis de donner une meilleure performance au système (visibilité et pertinence de résultats) par rapport à aux systèmes de référence, notamment le système traditionnel qui n'exploite aucune source d'enrichissement et le système monodimensionnel qui se base sur une seule vue de données pour décrire, enrichir et chercher l'information. Le modèle de recherche proposé inclut une technique d'interprétation de la requête utilisateur qui se base de son côté sur une représentation multidimensionnelle de son contenu. Cette représentation a permis de diversifier la recherche de l'utilisateur et a contribué à l'amélioration de la pertinence des résultats de recherche de l'utilisateur. Cette recherche est soutenue par une technique de désambiguïsation de la requête utilisateur lorsqu'un jeton de son contenu peut être lié à plusieurs interprétations. Cela a permis d'améliorer la précision du système en éliminant les documents qui peuvent ne pas répondre aux attentes de l'utilisateur.

2. Proposition d'un nouveau modèle générique de profil utilisateur. Afin de personnaliser les résultats de recherche de chaque utilisateur selon ses intérêts et ses préférences et améliorer davantage leur pertinence, un modèle de profil utilisateur est proposé. Il représente les données d'intérêt de l'utilisateur sous différents aspects exprimés en plusieurs niveaux d'abstraction. Cette représentation multi-niveaux a permis d'étendre la flexibilité des techniques qui sont appliquées pour intégrer le profil de l'utilisateur au sein du processus de RI. Elle a prouvé son efficacité au sein de plusieurs processus, notamment dans :

- La gestion de l'évolution des intérêts de l'utilisateur par la définition des sujets d'intérêt et des groupes de sujets fortement connexes. Cela a permis au système d'identifier le besoin informationnel de l'utilisateur lorsque sa requête de recherche est liée à plusieurs contextes de recherche.
- L'amélioration du processus de prédiction collaborative des intérêts de l'utilisateur en proposant un processus d'extraction des intérêts fréquents des utilisateurs à deux niveaux d'extraction, partant des sujets d'intérêt pour aller aux données d'interactions spécifiques.
- La définition de communautés d'intérêts à base de groupes de sujets fortement connexes. Ces communautés ont aidé à améliorer la recherche collaborative de l'utilisateur.

Plusieurs paramètres ont été intégrés pour améliorer l'exploitation de ce profil utilisateur, à savoir la fréquence des données, leur fraicheur et leur contexte. Ces paramètres ont aidé à définir les préférences de l'utilisateur en vue d'améliorer la prédiction de ses intérêts et l'identification de son besoin en information au cours d'une recherche ambiguë.

3. Recherche personnalisée de données. Afin d'enrichir la recherche de l'utilisateur, une nouvelle vue de données personnalisées est proposée à l'utilisateur. Cette vue de données exploite le profil de l'utilisateur pour personnaliser ses recherches selon ses intérêts et ceux de son voisinage. Ce profil a été exploité pour répondre à différents besoins :

- La désambiguïsation des requêtes ambiguës de l'utilisateur en identifiant le sujet de recherche qui couvre son besoin en information dans son profil.
- La désambiguïsation des objets de contenu (requêtes et étiquettes) au sein des règles d'association lors de l'inférence collaborative des intérêts.
- La description personnalisée des documents qui a permis de promouvoir ceux qui couvrent les intérêts récurrents de l'utilisateur exprimés par des requêtes/étiquettes.

- Le réordonnancement des résultats de recherche qui a aidé à privilégier les documents qui couvrent le sujet d'intérêt de l'utilisateur auquel est liée la requête de recherche.
- Enrichir davantage le contenu du profil utilisateur en l'intégrant dans un processus d'inférence collaborative d'intérêts.

4. Recommandation de données pour un nouvel utilisateur. Un modèle d'exploration et de recommandation de données est proposé pour un nouvel utilisateur. Il a pour objectif d'aider le système à construire le profil de cet utilisateur lorsqu'aucune information n'est disponible sur lui pour bénéficier de données de recommandation. Ce modèle est fondé sur l'organisation du contenu du système en communautés d'intérêts et le recours à des concepts de l'analyse des réseaux sociaux pour identifier les individus centraux au sein de chaque communauté. Cette proposition vise à faciliter l'exploration du contenu du système par les nouveaux utilisateurs.

5. Proposition d'un cadre d'évaluation du système. À cause de l'indisponibilité d'un cadre d'évaluation standard qui répond aux différents besoins des différents modèles proposés, un nouveau cadre d'évaluation est proposé. Il englobe plusieurs taches d'évaluations destinées à l'évaluation des modèles proposés, et propose des protocoles de construction de différentes collections de tests. Ce cadre pourra servir comme référence pour les systèmes basés sur les facettes de données et les systèmes de personnalisation orientée utilisateur. Les collections de tests qui sont proposées par ce cadre d'évaluation s'énumèrent comme suit :

- Une collection de données qui englobe les interactions réelles des utilisateurs sur un SRI basé sur les facettes de données (collection de requêtes annotées de sujets de recherche et de domaines de recherche, collection d'étiquettes, facettes d'intérêts, fichier de jugements de pertinence « requête-document», « étiquette-document », « facette-document »). La construction de cette collection est basée sur le développement d'un prototype fonctionnel d'un SRIF.
- Une collection de données étendue à base d'une collection sociale. Elle est fondée sur la création de requêtes de recherche simulées depuis les documents d'intérêts des utilisateurs.
- Une collection de requêtes ambiguës extraite depuis l'ontologie Dmoz. Elle inclut une collection de requêtes et des jugements de pertinence « requête-document ».

VIII.2. Limitations du système proposé

Malgré les résultats promoteurs qui sont obtenus dans ce travail de recherche en ce qui concerne les différents modèles proposés, il demeure des limitations que nous pouvons énumérer comme suit :

La première limitation concerne le modèle du profil utilisateur. La construction de ses niveaux de représentation abstraits dépend principalement du contenu textuel des ressources qui sont recueillies depuis les interactions de l'utilisateur, notamment les documents web. Ainsi, les améliorations qui ont été obtenues grâce à ces niveaux de représentation supérieurs ne peuvent être obtenues avec des systèmes qui proposent des informations multimédias telles que les images, les vidéos, les documents sonores, etc. L'association de ces entités à des concepts/catégories qui décrivent leur contenu s'avère un peu difficile. Plus concrètement :

- Comparativement à la solution apportée par le système de référence (Beldjoudi *et al.* 2017) pour la résolution de l'ambiguïté des étiquettes d'annotation associées aux documents qui se repose sur le calcul de similarités sociales entre les utilisateurs, notre proposition dépend principalement des classes sémantiques qui sont construites automatiquement au sein des profils des utilisateurs. Ainsi, lorsqu'il s'agit des étiquettes qui sont associées à des données multimédias, l'attribution de ces classes peut être difficile. Un autre point de comparaison qui peut être aussi citée et qui représente une limitation par notre contribution est qu'elle nécessite une phase d'apprentissage pour apprendre les profils des utilisateurs, ce qui n'est pas le cas de l'approche de (Beldjoudi *et al.* 2017) qui exploite directement les données qui sont fournies explicitement par les utilisateurs, à savoir, les étiquettes d'annotation des utilisateurs.
- Le modèle de représentation personnalisée des documents qui se repose sur le contexte des données pour sélectionner les données d'enrichissement pertinentes dépend aussi de la classe supérieure attribuée aux données d'intérêt des utilisateurs dans leurs profils. Ainsi, lorsque cette information est absente, l'amélioration n'est pas réalisable.
- La même chose vaut pour l'amélioration, en termes de temps et de pertinence des résultats, obtenue avec le processus d'extraction de corrélations entre les utilisateurs. Ce processus exploite les sujets d'intérêt des utilisateurs et les groupes de sujets connexes. Lorsque les ressources qui sont exploitées

par les utilisateurs à travers le système ne sont pas analysables, l'extraction de ces sujets représente une tâche plus complexe.

La deuxième limitation est liée au système d'extension QF qui s'occupe de créer les groupes de sujets fortement connexes des utilisateurs. Ces groupes sont obtenus à travers le calcul de similarités entre les graphes de sujets et la création de composantes connexe. Cependant, cette technique nécessite de nouveaux calculs lorsque de nouveaux sujets sont considérés par les utilisateurs et présente alors une limitation en ce qui concerne les calculs réguliers nécessaires.

VIII.3. Difficultés rencontrées

La mise en œuvre de nos modèles de recherche a été très couteuse en termes de temps et d'efforts. Les difficultés majeures qui sont liées à cette évaluation sont dues à l'indisponibilité de collections de données adaptables aux besoins d'un système multidimensionnel et hybride comme le nôtre, et à la difficulté de créer des collections qui répondent aux interactions possibles de l'utilisateur avec de tels systèmes et surtout l'incapacité de couvrir toutes les caractéristiques possibles.

C'est la raison pour laquelle une évaluation hybride a été adoptée. Elle combine deux techniques d'évaluation. La première évaluation fait appel à de vrais utilisateurs, et la deuxième se base sur la création d'utilisateurs simulés. La première évaluation a permis d'évaluer l'efficacité du système multi-facettes, mais à cause de son petit volume, elle n'a malheureusement pas pu être exploitée dans le reste des expérimentations. Une nouvelle collection de tests a été donc créée. Elle est fondée sur l'exploitation de la collection sociale Delicious pour simuler des requêtes de recherche à partir des documents d'intérêt des utilisateurs. La nouvelle collection étendue reste aussi réelle puisqu'elle se base sur des interactions réelles effectuées par de vrais utilisateurs sur le réseau social précité. Le but est d'étendre son contenu avec une collection de requêtes de recherche. Cette extension a permis de tester l'efficacité des requêtes dans l'amélioration d'extraction de corrélations entre les utilisateurs et de démontrer l'insuffisance des étiquettes dans cette tâche d'extraction. En effet, la combinaison de ces requêtes aux étiquettes a permis d'améliorer cette corrélation en augmentant la présence simultanée des intérêts dans les profils des utilisateurs.

VIII.4. Futurs travaux

Cette thèse ouvre la voie à plusieurs perspectives qui sont liées au même contexte des problèmes abordés.

- Développer une approche pour l'inférence automatique des espaces de projections à partir d'un corpus de documents. Mettre en production l'implémentation de notre SRI après son amélioration (indexation à grande échelle des documents web, amélioration de l'interface graphique, etc.). Puis évaluer la nouvelle version pour valider les conclusions obtenues.
- Évaluer notre système sur d'autres collections de tests plus volumineuses pour tester le passage à l'échelle de ce système.
- Il sera intéressant d'étudier d'autres catégories d'informations sociales à exploiter pour identifier les intérêts de l'utilisateur et son voisinage. Par exemple, les interactions entre les utilisateurs, l'échange de données, échanges de message, etc. Il sera intéressant aussi d'étudier d'autres informations contextuelles qui aident à mieux comprendre le comportement de l'utilisateur et améliorer davantage ses recherches, telles que les évènements qui sont liés à ces recherches.
- Élargissement du profil utilisateur à d'autres types de données (ex: données multimédias telles que les vidéo, images, etc.).
- Proposer une technique d'indexation des documents à base de profils des utilisateurs qui soit moins couteuse en mémoire, ceci est possible par l'utilisation de techniques de regroupement pour retrouver les profils types.
- Étendre la personnalisation des données inter-espace.
- Les techniques de personnalisation de données proposées dans cette thèse peuvent être testées dans la personnalisation de l'accès aux services web dans une application orientée service. Elles peuvent être utiles pour les fournisseurs de services pour l'organisation de leurs services dans l'annuaire. Ainsi, les services peuvent être répartis en plusieurs groupes définis selon les catégories des services les plus connexes, et la découverte de ces services peut être effectuée en identifiant le groupe d'intérêt de l'utilisateur le plus représentatif de ses besoins puis la sélection du groupe de services qui correspond à ce groupe d'intérêt. Cela aide à optimiser l'accès à ces services et promet une sélection pertinente qui répond aux attentes de chaque utilisateur.

- La prédiction qui se base sur l'identification des préférences des utilisateurs peut être utile pour les sites e-commerce en leur permettant de recommander aux consommateurs des articles moins nombreux et de meilleure qualité, surtout que ces consommateurs dépendent aujourd'hui de plus en plus de leurs appareils mobiles dont l'écran est petit et n'accepte pas une grande quantité de données. Cette prédiction à base de préférences aide à réduire la quantité des recommandations tout en s'assurant de la pertinence des résultats.
- Tester avec de vrais utilisateurs les filtres temporels qui sont proposés dans le chapitre 4 (cf. tableau 4.3). Ces filtres permettent à l'utilisateur d'exprimer ses préférences pour un sous-ensemble de données d'intérêt au sein de son profil. Les résultats personnalisés qui peuvent être obtenus avec ces filtres sont utiles pour tester avec de vrais utilisateurs l'efficacité de notre technique de détection automatique des préférences utilisateurs proposée dans le chapitre 5. Cela est possible en comparons les résultats de la technique automatique avec ceux de la technique interactive à base de filtres où l'utilisateur exprime explicitement ses préférences en appliquant un filtre temporel. Les deux méthodes, automatique et interactive, peuvent être combinées au sein du système pour améliorer le processus de RI.
- La possibilité d'intégrer les différentes techniques proposées à travers des formalismes unificateurs.
- Le modèle de recommandation proposé dans cette thèse pour assister un nouvel utilisateur dans ses premières recherches soulève des perspectives intéressantes. L'expérimentation de ce modèle n'a pas été effectuée dans ce travail. Il sera intéressant de pouvoir tester et évaluer son efficacité dans de futurs travaux. Le problème de démarrage à froid est très fréquent dans la recommandation e-commerce lorsque peu, voir pas, d'information sur un utilisateur n'est connu par le système. La résolution de ce problème aide l'utilisateur à recevoir rapidement des recommandations intéressantes qui répondent à ses attentes.
- Les contributions de cette thèse qui ne sont pas encore publiées vont être prochainement soumises dans des journaux scientifiques internationaux. Le protocole d'évaluation proposé dans le chapitre 7 va être peaufiné pour être diffusé au sein d'une revue scientifique.
- Nous envisageons de mettre à disposition de la communauté de la RI, les collections de tests construites. Elles seront préparées sous une forme plus compréhensible et exploitable telles que les formes utilisées par les collections de tests standards de TREC.

ANNEXES

ANNEXE 1 : Processus de construction des clusters de requêtes de recherche à base de sujets de recherche

Entrée :
q : requête de recherche
Ω: seuil de corrélation
O_D: ontologie topique
cl_i : un cluster de requêtes
$V_A.cl_i$: un cluster de requêtes lié à une activité de recherche

Sortie :
Prs : Ensemble de profils d'activités
E_{CL}: Ensemble de clusters de requêtes

```
1.  /- compteur de clusters
2.  i = 1;
3.  /- Initialiser le cluster
4.  Pour chaque nouvelle requête q Faire :
5.   /-Collecte de documents pertinents derrière q
6.       D_q = (d1,....., dn)
7.  /- Construire le profil d'activité V_A^q
8.     /- Projection de D_q sur O_D
9.       V_A^q ← P_D^ODP = ((c1,score1),..., (ck,scorek) )
10. /- tester si l'ensemble des clusters est vide
11.    Si E_CL = ∅ alors
12.     /- il s'agit de la première requête
13.     /- Ajouter le profil de l'activité courante dans l'ensemble des profils d'activités
14.      Prs ← V_A^q
15.     /- Initialiser le premier cluster avec la première requête
16.       cl_1 ← q
17.     /- Ajouter le premier cluster dans l'ensemble des clusters E_CL
18.      E_CL ← cli
19.    /- Sinon si E_CL ≠ ∅ alors
20.     Pour chaque profil d'activité V_A dans l'ensemble Prs Faire
21.     /- Evaluer la corrélation entre le profil de l'activité courante V_A^q et le profil V_A
22.          Corr = CorrKendall (V_A^q, V_A)
23.      Si Corr > Ω Faire
24.       /- Mettre à jour le profil V_A avec le contenu du profil courant V_A^q
25.          V_A= V_A^q+ V_A
26.       /- Ajouter la requête q dans le cluster correspondant au profil corrélé V_A
27.          V_A.cl ← q
28.      Sinon
29.          /- Ajouter V_A^q dans l'ensemble Prs
30.           Prs ← V_A^q
31.         /- Ajouter la requête dans un nouveau cluster
32.            i = i++
33.           cl_i ← q
34.         /- Ajouter le cluster dans l'ensemble
35.           E_CL← Cli
36.        Fin Sinon
37.    Fin Sinon
38. Fin
```

Algorithme 1 : Processus de construction des clusters de requêtes à base de sujets de recherche

ANNEXE 2 : Algorithme de sélection des objets d'enrichissement d'un document à base d'une pertinence hybride de contenu

Entrée: d_k : document cible
U: les utilisateurs du système
$OB_{d_k}^{u_j} = \{Q_{u_j} \cup T_{u_j}\}$: liste d'objets d'enrichissement de départ de l'utilisateur u_j pour d_k
SRIF: système de recherche d'information multi-facettes
D_{obj_i}: Top k-documents retournés derrière une recherche avec obj_i
$Rep_{obj_i}^{d_k}$: représentativité d'un objet d'enrichissement obj_i pour d_k
QF: système d'analyse de données
Sortie : OB'_{d_k}: ensemble d'objets d'enrichissement retenus pour d_k

Début

1. $Rep_{obj_i}^{d_k}$ = Faux;
2. D_{obj_i} =∅ ;
3. OB'_{d_k} =∅ ;
4. Pour chaque $u_j \in$ U faire
5. **Pour** chaque $obj_i \in OB_{d_k}^{u_j}$ **faire**
6. **Si** $\exists\, u_j \in U: < obj_i, u_j, d_k > \in R$ **alors**
7. $Rep_{obj_i}^{d_k} = Vrai$
8. $OB' = OB'\ U\ obj_i$
9. **Sinon** // interroger SRIF avec obj_i:
10. **Si** $obj_i \in Q_{d_k}$ **alors**
11. $D_{obj_i} \leftarrow$ InterrogerSRIF($obj_i, EspaceIdentitaire$, *EspaceSémantique*);
12. **Sinon si** $obj_i \in T_{d_k}$ **alors**
13. $D_{obj_i} \leftarrow$ InterrogerSRIF($obj_i, EspaceSocial$);
14. **Si** $d_k \in D_{obj_i}$ **alors**
15. $Rep_{obj_i}^{d_k} = Vrai$
16. $OB' = OB'U\ obj_i$
17. **Sinon** rejeter (obj_i);
18. **Fin Pour;**
19. **Fin Pour;**
20. **Retourner** OB'_{d_k}**;**

Fin

Algorithme 2. Algorithme de sélection des objets d'enrichissement pour un document d_k à base de la représentativité du contenu

ANNEXE 3 : Processus de construction du profil de la requête (conceptualisation)

La conceptualisation d'une requête de recherche consiste à associer à son contenu un ensemble de concepts qui correspondent à des domaines d'intérêt/catégories dans une ontologie de référence. Cela est effectué en projetant le contenu de la requête sur le contenu de l'ontologie. Le résultat est un ensemble de concepts pondérés avec des scores de similarité sémantique qui traduisent chacun un degré de correspondance de la requête avec l'ensemble des domaines obtenus. Afin de sélectionner les concepts les plus pertinents pour cette requête, les concepts obtenus sont représentés sous un graphe conceptuel noté par G, tel que G = (E, V) où E est un ensemble des concepts et V est l'ensemble des relations

sémantiques qui les relient les uns aux autres. Les concepts qui restent déconnectés sont considérés comme non pertinents et sont donc éliminés.

La connexion entre les concepts obtenus consiste à chercher dans l'ontologie de référence (Dmoz dans notre cas) s'il existe un chemin P entre les couples de concepts (c_i, c_j). Un chemin P (i, j) liant deux concepts c_i et c_j peut être un lien direct e_{ij} ou un lien indirect passant par d'autres concepts c_k, tel que P (i, j) = {e_{in},., e_{kj}}. Un concept c_k peut être un membre de la liste initiale des concepts comme il peut être un nouveau concept. Nous appelons les nouveaux concepts par les concepts de liaison. Ils servent d'un côté à relier entre deux concepts initiaux lorsqu'il n'existe aucun lien direct ou un chemin passant par d'autres concepts initiaux entre eux, et d'un autre côté de diversifier la recherche en étendant la liste initiale des concepts. L'objectif est de construire un ou plusieurs graphes reflétant le profil de la requête. Cette multivalence de graphes s'explique par le faite qu'une requête peut être liée à plusieurs sujets de recherche. Chaque sujet de recherche est représenté par un graphe. Ces graphes sont déconnectés les uns aux autres.

Pour relier les concepts entre eux, nous nous basons sur une méthode de découverte de concepts qui identifie pour chaque concept ses concepts fils dans l'ontologie de référence (Daoud *et al.* 2010b). Puis, appliquer une technique de diffusion de poids qui permet d'ajuster les concepts initiaux avec de nouveaux poids et d'attribuer des poids aux nouveaux concepts selon les liens sémantiques qui les relient aux autres concepts (Daoud *et al.* 2010b). Pour ce faire, nous exploitons la technique de pondération des liens définie dans les travaux de (Maguitman *et al.* 2005). Cette technique de pondération attribue pour chaque lien de l'ontologie Dmoz une pondération. Ainsi, le score attribué à chaque concept différera selon les types de relations qui le relient aux autres concepts. Chaque concept C_i appartenant à la liste initiale diffuse son poids à ses nœuds voisin c_j en se basant sur les types de relations qui les relient ensemble. Dans cette ontologie, un concept peut avoir plusieurs pères, cela est possible à travers les relations non hiérarchiques (S et R). Par conséquent, un concept peut être activé par plusieurs concepts durant le processus de diffusion de poids. Le score final est la moyenne des scores qui lui sont diffusés. Selon les auteurs (Maguitman *et al.* 2005), lorsqu'un chemin entre deux concepts c_i et c_j contient plus d'une relation de type S or R, ces deux concepts ne sont pas considérés de la même famille. Nous prenons en considération cette information durant le processus de diffusion de poids. Ainsi, un concept C_i ne

transmet le poids qu'il a eu d'un autre concept c_j aux autres concepts c_k que lorsqu'ils font partie de la même famille.

ANNEXE 4 : Démonstration théorique du processus d'extraction des itemsets fréquents selon le modèle classique et selon notre modèle

Exemple 1. Considérons l'exemple d'une base d'activités de 13 données d'intérêt relatives à 10 utilisateurs. Ces données peuvent représenter des requêtes de recherche ou des étiquettes. Elles appartiennent à 4 sujets différents Suj_1, Suj_2, Suj_3 et Suj_4, tel qu'illustré dans la partie-1 du tableau 1.

Suj_1 = {K, P, R, S}

Suj_2 = {A, B}

Suj_3= {C, D, E, F}

Suj_4 ← {L, M, N}

Partie-1

Utilisateur	Transaction basée sur les données granulaires	Transaction basée sur les sujets de recherche
u_1	PKCDFE	Suj1, Suj3
u_2	AML	Suj2, Suj4
u_3	SKNM	Suj1, Suj4
u_4	CDE	Suj3
u_5	CDF	Suj3
u_6	CDFL	Suj3, Suj4
u_7	EL	Suj3, Suj4
u_8	L	Suj4
u_9	BM	Suj2, Suj4
u_{10}	M	Suj4

Partie-2

Tableau 1. Exemple 1 de données transactionnelles des utilisateurs

Si on suppose que le support minimum pour considérer un itemset fréquent est de 0.3. Nous présentons une comparaison entre le processus classique qui se base pour l'extraction des itemsets fréquents sur la préparation des données transactionnelles des utilisateurs sous une forme granulaire, et notre modèle qui se base sur une extraction à deux niveaux allant des sujets fréquents jusqu'aux données granulaires fréquentes. La partie 2 du tableau 1 présente les deux formes de transactions. La première est une représentation granulaire des activités utilisateurs sur laquelle se base le modèle classique, et la deuxième est la représentation générique des données d'intérêt à base de sujets d'intérêt sur laquelle se base notre modèle pour préparer les données de départ. Durant tout le processus d'extraction la règle suivante est appliquée au sein des deux modèles : R1= « si un itemset n'est pas fréquent, tous ses supersets ne le seront pas non plus, donc ils sont éliminés ». C'est la règle sur laquelle se base l'algorithme Apriori pour l'extraction des itemsets fréquents (Agrawal *et al.* 1993) (cf. figure 1). Nous commençons par donner un rappel sur quelques notions de base exploitées dans ces exemples.

Définition 1. Itemsets. Un itemset est un ensemble d'items. Un item peut représenter n'importe quel objet, dans notre cas il peut être une donnée granulaire (étiquette ou requête) ou une donnée générique (un sujet d'intérêt). Donc, un itemset spécifique est un ensemble d'items granulaires (ensemble d'étiquettes ou de requêtes), un itemset générique est un ensemble de sujets. Un k-itemset (spécifique/générique) est un ensemble de k items (spécifiques ou génériques).

Définition 2. **Superset**. Un superset de A est un itemset descendant de A dans le treillis des items (cf. figure 1). On dit que B est un superset de A si card(A) < card(B) et A $\subset$ B => sup(B) $\leq$ sup(A). Par exemple dans la figure 1, l'itemset générique S1S2 est un superset de S1 et aussi superset de S2.

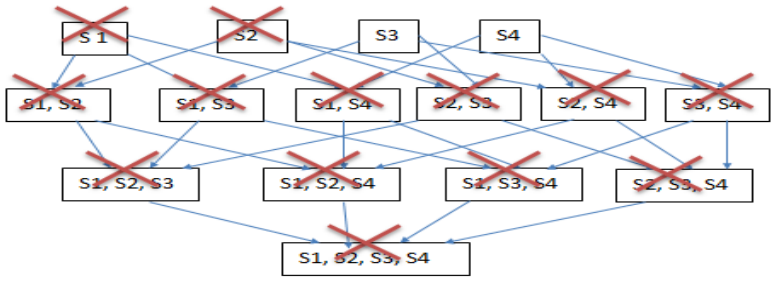


Figure 1. Exemple applicatif de la règle R1 sur les données d'activités génériques des utilisateurs

Cas 1 -Extraction classique : extraction des itemsets spécifiques fréquents.

Étape 1-Calcul des supports des 1-itemsets spécifiques

1-itemset	support	Support min=0.3
A	0.1	éliminé
B	0.1	éliminé
C	0.4	retenu
D	0.4	retenu
E	0.3	retenu
F	0.3	retenu

1-itemset	support	Support min=0.3
K	0.2	éliminé
P	0.1	éliminé
R	0	éliminé
S	0.1	éliminé
L	0.4	retenu
M	0.4	retenu
N	0.1	éliminé

⇨ **#calculs =13*10**

Tableau 2. Valeurs des supports des 1-itemsets spécifiques

En appliquant la règle R1, les supersets de A, B, K, P, R, S, N sont éliminés. Ainsi, l'étape 2 représente les 2-itemsets qui sont construits à base des 1-itemsets retenus de l'étape précédente, à savoir C, D, E, F, L, M.

Étape 2- Calcul des supports des 2-itemsets spécifiques

2-itemset	support	Support min=0.3
CD	0.4	retenu
CE	0.2	éliminé
CF	0.3	retenu
CL	0.1	éliminé
CM	0	éliminé
DE	0.2	éliminé
DF	0.3	retenu

2-itemset	support	Support min=0.3
DL	0.1	éliminé
DM	0	éliminé
EF	0.1	éliminé
EL	0.1	éliminé
EM	0	éliminé
FL	0.1	éliminé
FM	0	éliminé
LM	0.1	éliminé

⇨ **#calculs =15*10**

Tableau 3. Valeurs des supports des 2-itemsets spécifiques

Les 2-temsets qui sont retenus de cette étape sont CD, CF. DF. Jusque-là, le nombre d'opérations effectuées est de 28 opérations, en se basant sur le même principe, le reste des étapes est effectué jusqu'à finir avec les k-itemsets possibles. Le but de cette démonstration est de comparer le nombre d'opérations effectuées par les deux modèles.

Nous passons maintenant aux étapes illustratives de notre modèle et calculons le nombre d'opérations nécessaires pour arriver à cette étape d'extraction, c'est-à-dire l'obtention de l'ensemble {CD, CF. DF}.

Cas 2 : Extraction d'itemsets fréquents à deux niveaux d'extraction (notre modèle)

Étapes 1 : Calcul des supports des 1-itemsets et 2-itemsets génériques

1-itemset	support	Support min=0.3
Suj1	0.2	éliminé
Suj2	0.2	éliminé
Suj3	0.5	retenu
Suj4	0.6	retenu

Partie-1

2-itemset	support	Support min=0.3
Suj3 Suj4	0.2	éliminé

⇨ **#calculs =5*10**

Partie-2

Tableau 4. Valeurs des supports des 1-itemsets et 2-itemsets génériques

En se référant à la partie 1 du tableau 4, les 1-itemsets qui sont retenus sont Suj_3 et Suj_4. En suivant la règle R1, les 2-itemsets sont construits uniquement à partir de cet ensemble. Le seul 2-itemset générique qui peut être construit est Suj_3Suj_4. Comme on peut le voir à travers la partie 2 du tableau 4, le support de cet itemset est inférieur au seuil qui a été défini, il donc est illimité. Ainsi, le système se base sur les deux items génériques Suj3 et Suj4 pour extraire les k-itemsets spécifiques fréquents. Ceux-ci ont des supports de 0.5 et 0.6. Il s'agit des k-itemsets spécifiques extraits des deux ensembles {C, D, E, F} et {L, M, N}. Pour ce faire, les deux ensembles Suj_3 et Suj_4 sont délimités puisque l'itemset Suj_3Suj_4 n'est pas fréquent donc aucun k-itemset spécifique constitué d'une combinaison de leurs items granulaires n'est fréquent. C'est la règle sur laquelle se base notre modèle. Cette règle est comme suit : « R2 = si un itemset

générique (Suj_i, Suj_k) n'est pas fréquent, les itemsets spécifiques construits par combinaison de leurs items spécifiques ne sont pas également fréquents ». Cela peut être vérifié à travers le tableau 3 en haut et la figure 2 ci-dessous. On peut bien voir que tous les 2-itemsets spécifiques constitués par combinaison d'un élément de Suj3 et d'un autre de Suj4 ont un support inférieur au seuil défini.

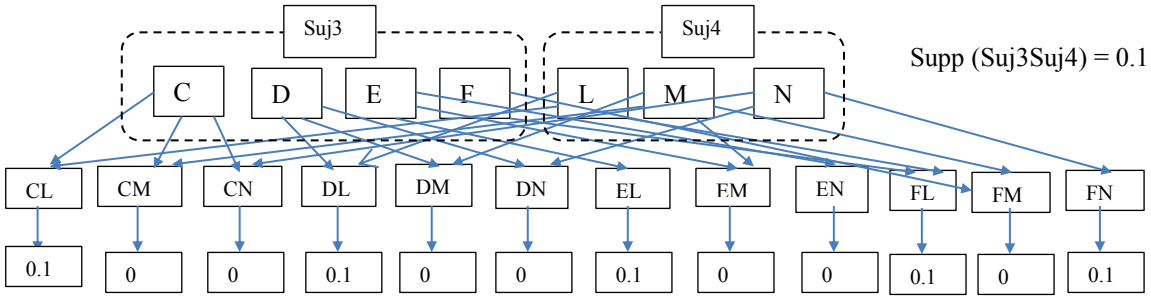


Figure 2. Les valeurs des supports des 2-itemsets spécifiques construits par combinaison d'items spécifiques des deux itemsets génériques Suj3 et Suj4.

Étape 2- extraction des 1-itemsets et 2-itemsets spécifiques fréquents

Étape 2.1- extraction des 1-itemsets et 2-itemsets spécifiques fréquents à partir de Suj3 = {C, D, E, F}

1-itemset	support	Support min=0.3
C	0.4	retenu
D	0.4	retenu
E	0.3	retenu
F	0.3	retenu

2-itemset	support	Support min=0.3
CD	0.4	retenu
CE	0.2	éliminé
CF	0.3	retenu
DE	0.2	éliminé
DF	0.3	retenu
EF	0.1	éliminé

⇨ **#calculs =10*10**

Tableau 5. Valeurs des supports des 1-itemsets et 2-itemsets spécifiques

Étape 2.2- extraction des 1-itemsets et 2-itemsets spécifiques fréquents à partir de Suj4 = {L, M, N}

1-itemset	support	Support min=0.3
L	0.4	retenu
M	0.4	retenu
N	0.1	éliminé

2-itemset	support	Support min=0.3
LM	0.1	éliminé

⇨ **#calculs =4*10**

Tableau 6. Valeurs de supports des 1-itemsets et 2-itemsets spécifiques

Les 2-itemsets spécifiques retenus de cette étape sont CD, DF, EF. Le reste des étapes présentent le même nombre d'opérations pour les deux modèles que nous notons par n_1, sachant que 10 transactions

existent (nombre d'utilisateurs), les nombres d'opérations effectuées par notre modèle et le modèle classique sont de respectivement de $19*10+n_1$ et $28*10+n_1$ opérations.

Nous pouvons à travers cet exemple que i) les sujets fréquents englobent des objets granulaires non fréquents », ii) lorsque les k-itemsets génériques sont non fréquents, les k-itemsets obtenus par combinaison de leurs items granulaires sont non fréquents.

Exemple 2. Dans ce deuxième exemple, nous présentons le cas où les utilisateurs partagent les mêmes sujets d'intérêt, mais ne possèdent pas les mêmes données d'activités granulaires. Nous traitons à travers cet exemple le cas où un itemset générique est fréquent tandis que ses itemsets spécifiques ne le sont pas. Nous analysons ainsi l'influence de notre modèle sur le temps de calcul des itemsets spécifiques fréquents par rapport au modèle classique. Ce temps est estimé en termes de nombre d'opérations effectuées pour l'extraction de ces itemsets fréquents.

Partie-1
$Suj_1 \leftarrow$ K, P, R, S
$Suj_2 \leftarrow$ A, B
$Suj_3 \leftarrow$ C, D, E, F
$Suj_4 \leftarrow$ L, M, N

Partie-2

Utilisateur	Transaction basée sur les données granulaires	Transaction basée sur les sujets de recherche
u_1	KPM	Suj1, Suj4
u_2	AB	Suj2
u_3	RD	Suj1, Suj3
u_4	RB	Suj2
u_5	CF	Suj3
u_6	CFM	Suj3, Suj4
u_7	SKAB	Suj1, Suj2

Tableau 7. Exemple 2 de données transactionnelles des utilisateurs

Cas 1. Modèle classique

1-itemset	support	Support min=0.3
A	0.28	éliminé
B	0.42	retenu
C	0.28	éliminé
D	0.14	éliminé
E	0	éliminé
F	0	éliminé

1-itemset	support	Support min=0.3
K	0.14	éliminé
P	0.14	éliminé
R	0.28	éliminé
S	0.14	éliminé
L	0	éliminé
M	0.28	éliminé
N	0	éliminé

⇨ **#calculs =13*7**

Tableau 8. Les valeurs des supports des 1-itemsets spécifiques

Le seul 1-itemsets retenu est l'itemsets B. Fin du processus.

Cas 1. Notre modèle

1-itemset	support	Support min=0.3
Suj1	0.42	retenu
Suj2	0.42	retenu
Suj3	0.42	retenu
Suj4	0.28	éliminé

2-itemset	support	Support min=0.3
Suj_1Suj_2	0.14	éliminé
Suj_1Suj_3	0.14	éliminé
Suj_2Suj_3	0	éliminé

⇨ **#calculs =7*7**

Tableau 9. Les valeurs des supports des 1-itemsets et 2-itemsets génériques

Les itemsets génériques qui sont retenus de cette étape sont Suj1, Suj2, et Suj3. En appliquant la règle R2, les itemsets spécifiques vont être extraits en délimitant les items spécifiques relatifs aux items génériques, car tous les 2-itemsets génériques sont éliminés (leurs supports sont inférieurs au seuil min) (cf. figure 3).

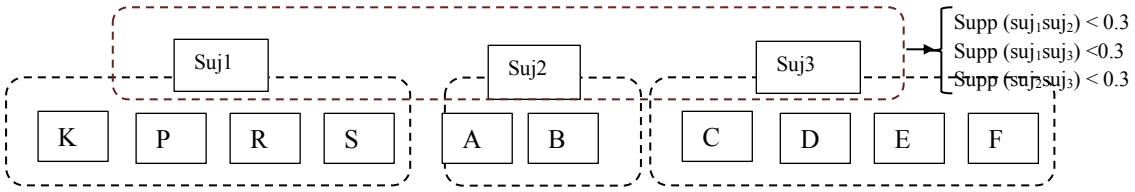


Figure 3. Items spécifiques délimités selon la règle R2

1-itemset	support	Support min=0.3
A	0.28	éliminé
B	0.42	retenu
C	0.28	éliminé
D	0.14	éliminé
E	0	éliminé
F	0	éliminé

1-itemset	support	Support min=0.3
K	0.14	éliminé
P	0.14	éliminé
R	0.28	éliminé
S	0.14	éliminé

⇨ **#calculs =10*7**

Le seul 1-itemset spécifique retenu de cette étape est B.

Nous pouvons voir que le nombre d'opérations considérées par le modèle classique pour l'extraction des itemsets spécifiques fréquents est de 91 opérations (13*7), il est inférieur au nombre d'opérations effectuées par le modèle classique qui s'élève à 119 opérations (17*7). Ainsi, dans de tels cas les opérations effectuées sur les données génériques peuvent avoir un impact négatif sur le temps de calcul.

ANNEXE 5 : Extension de la collection de tests Delicious par simulation de requêtes de recherche

Objectif. Ce processus consiste à étendre la collection de tests extraite à travers notre système avec d'autres utilisateurs simulés. Cette simulation aide à enrichir la base des profils utilisateurs au sein de notre collection de tests avec d'autres utilisateurs.

Motivation. Cette extension se base sur les deux motivations suivantes :

- Le modèle de recommandation de données proposé pour l'enrichissement du profil utilisateur se base sur l'analyse de corrélations entre les comportements des utilisateurs à travers le système. Cela nécessite une grande collection de tests avec une grande corrélation entre les activités de ces utilisateurs. Malheureusement, la base des profils collectée à travers notre système est considérée comme petite et insuffisante pour de tels tests. Il est donc primordial d'augmenter son contenu avec d'autres utilisateurs simulés. Cela se base sur l'exploitation de la collection sociale extraite grâce à notre explorateur de données que nous avons nommé par la collection de départ (cf. tableau 7.1) et la proposition d'une stratégie de simulation de requêtes qui permettent d'enrichir les profils sociaux de cette collection avec des requêtes de recherche. La construction de ces requêtes permet d'analyser leur efficacité pour la détection de corrélation entre les utilisateurs. Ces requêtes peuvent contribuer à l'augmentation de corrélations entre les activités de ces utilisateurs, d'autant plus que les étiquettes qui sont employées par ces derniers sont généralement personnelles, ambiguës et imprécises (cf. section 1.1.7). Ainsi, l'exploitation seule de ces étiquettes réduit la possibilité de détecter des corrélations intéressantes pour la recommandation collaborative.
- L'évaluation à base d'utilisateurs réels est généralement efficace en termes d'utilité et d'utilisabilité réelles. Elle n'est malheureusement pas toujours faisable, et surtout demande beaucoup de temps comparativement à l'évaluation par simulation d'utilisateurs. Le recours à une combinaison des deux techniques précitées est nécessaire pour l'obtention d'une évaluation qui soit la plus juste possible (Daoud *et al.* 2010b).

Stratégie de construction. Il consiste à utiliser la collection de tests de départ extraite avec le module d'extraction de données (cf. section VII.2.1) pour construire des requêtes de recherche simulées. Cette collection de départ comporte une collection de documents, une collection d'étiquettes et les profils des

utilisateurs qui englobent des listes de jugements de pertinence associant pour chaque utilisateur les documents pertinents à ses étiquettes d'annotation (cf. tableau 7.1).

Dans une recherche full texte, une requête de recherche est constituée d'un ensemble de jetons que le SRI utilise pour localiser des documents dans l'index. Des études ont été effectuées dans la recherche des documents structurés (Hu *et al.* 2005) (Zhai 2008). Elles ont montré que les utilisateurs ont tendance à cibler leur recherche sur les titres des documents, et que lorsque ces titres sont pris en compte dans la recherche, les résultats sont plus précis. De là, nous pouvons déduire que les requêtes utilisées par les utilisateurs pour exprimer leurs besoins peuvent correspondre aux titres des documents recherchés. C'est sur la base de cette analyse que nous proposons de simuler nos requêtes à partir des titres des documents.

Ce processus de simulation de requêtes est illustré dans la figure 4 et se résume par les deux étapes suivantes :

- Extraction des activités de recherche à base de fréquences simultanées d'étiquettes. Dans la collection Delicious, chaque utilisateur est lié à un ensemble d'étiquettes qu'il a utilisé. Chaque étiquette est liée à son tour à un ensemble de documents. Si nous considérons chaque étiquette et ses documents pertinents comme une seule activité de recherche cela risque d'augmenter le nombre de calculs qui sont liés au modèle de construction du profil multi-niveaux (cf. figure 7.11). La détection des activités similaires a pour objectif d'extraire le maximum de documents qui peuvent représenter une seule activité. Cette étape se base sur le calcul des fréquences simultanées des étiquettes dans les profils des utilisateurs (cf. figure 4, partie 1). Pour chaque utilisateur, un ensemble d'activités est extrait. Chacune est représentée par une requête simulée à partir de l'ensemble des étiquettes qui apparaissent simultanément et fréquemment dans son profil. Chaque activité est associée à un ensemble de documents pertinents (cf. figure 4, partie 2).
- Une fois que les activités sont extraites, des requêtes sont simulées depuis les titres des documents obtenus au sein des activités simulées (cf. figure 4, partie 3). La construction de ces requêtes se base sur le calcul des fréquences simultanées des jetons dans les titres de ces documents.

- Le résultat final est un ensemble d'activités associé à chaque utilisateur. Chaque activité est représentée par une requête de recherche qui lui correspond un ensemble de documents pertinents. Chaque document est annoté avec un ensemble d'étiquettes.

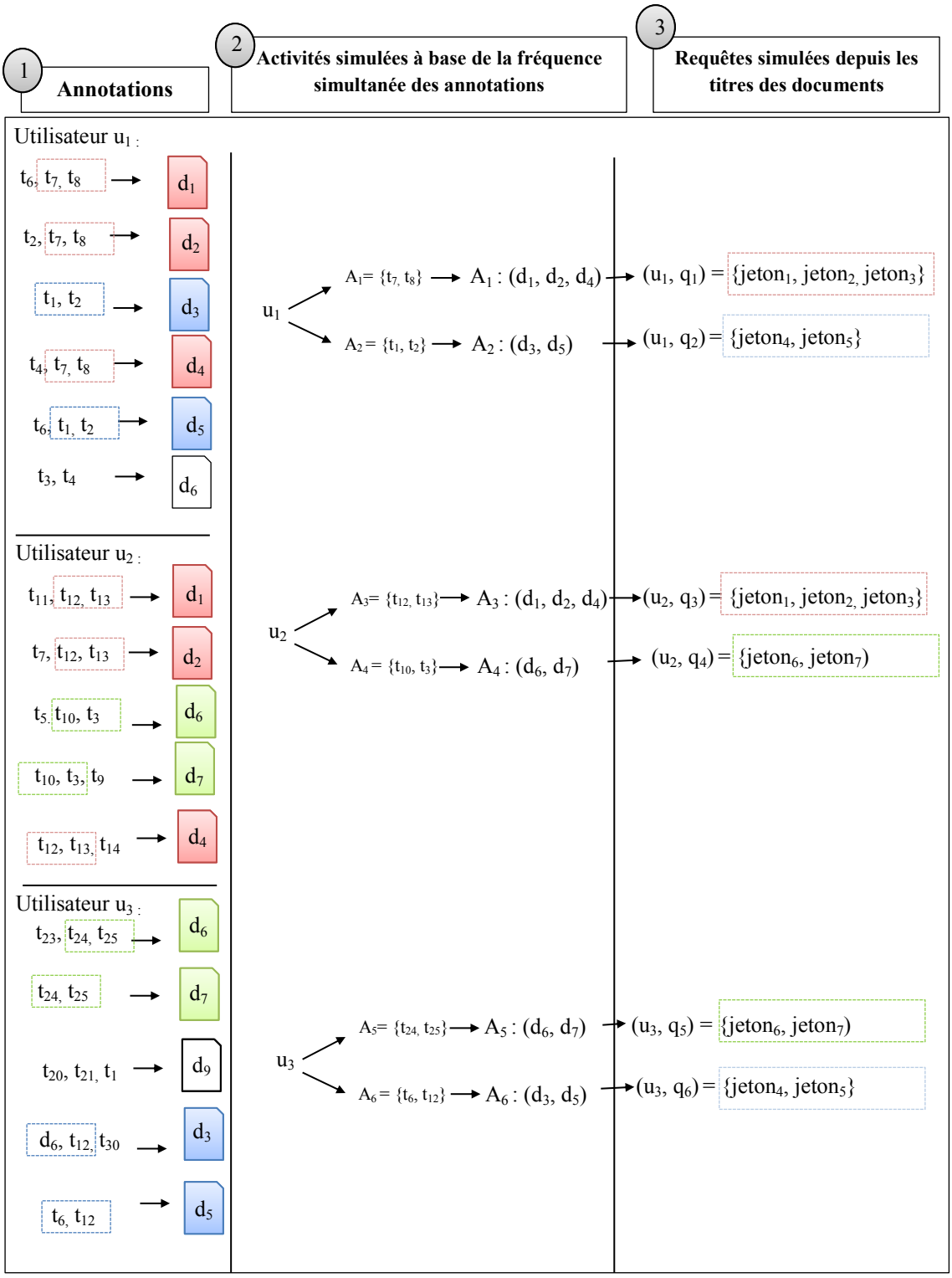


Figure 4. Processus de simulation de requêtes à partir d'une collection de données sociale

Comme nous pouvons le voir les utilisateurs qui partagent des documents d'intérêt en commun les ont annotés différemment (cf. figure 4 partie 1), cela n'aide pas le système à extraire les corrélations entre ces utilisateurs. Lorsque les requêtes de recherche sont extraites depuis les titres des documents cela a permis d'extraire les corrélations entre ces utilisateurs (cf. figure 4, partie 3).

Bibliographie

Abdel-Hafez A et Xu Y. 2013. A survey of user modelling in social media websites. Computer and Information Science, 6 : 59.

Abdulahhad K, Chevallet J-P et Berrut C. Solving concept mismatch through bayesian framework by extending umls meta-thesaurus. Dans : CORIA 2011-COnférence en Recherche d'Information et Applications, 2011. Editions Universitaires d'Avignon, p. 311-326.

Abel F, Gao Q, Houben G-J et Tao K. Semantic enrichment of twitter posts for user profile construction on the social web. Dans : Extended Semantic Web Conference, 2011. Springer, p. 375-389.

Achemoukh F et Ahmed-Ouamer R. Representation and Evolution of User Profile in Information Retrieval Based on Bayesian Approach. Dans : International Symposium on Methodologies for Intelligent Systems, 2014. Springer, p. 486-492.

Adda M. 2008. Intégration des connaissances ontologiques dans la fouille de motifs séquentiels avec application à la personnalisation Web. Université des Sciences et Technologie de Lille-Lille I; Université de Montréal.

Adda M, Missaoui R et Valtchev P. 2007. Relation rule mining. The International Journal of Parallel, Emergent and Distributed Systems, 22 : 439-449.

Adda M, Hannech A et Mcheick H. New web information retrieval paradigm based on a multi-space interpretation index and projection operations. Dans : Information, Communication and Automation Technologies (ICAT), 2013 XXIV International Symposium on, 2013. IEEE, p. 1-7.

Agrawal R, Imieliński T et Swami A. Mining association rules between sets of items in large databases. Dans : Acm sigmod record, 1993. ACM, p. 207-216.

Aicha Aggoune AB, Mohamed Khiereddine Kholladi. 2016. Enhancement of Indexing Model for Heterogeneous Multimedia Documents: User Profile Based Approach. International Journal of Business and Economics Engineering, 3.

Allan J, Aslam J, Belkin N, Buckley C, Callan J, Croft B, Dumais S, Fuhr N, Harman D et Harper DJ. Challenges in information retrieval and language modeling: report of a workshop held at the center for intelligent information retrieval, University of Massachusetts Amherst, September 2002. Dans : ACM SIGIR Forum, 2003. ACM, p. 31-47.

Almazro D, Shahatah G, Albdulkarim L, Kherees M, Martinez R et Nzoukou W. 2010. A survey paper on recommender systems. arXiv preprint arXiv:10065278.

Alonso O, Gertz M et Baeza-Yates R. On the value of temporal information in information retrieval. Dans : ACM SIGIR Forum, 2007. ACM, p. 35-41.

Alsalama A. 2013. A Hybrid Recommendation System Based on Association Rules.

Amar M. 2009. Taxonomies, ontologies et folksonomies :situer les différences et identifier les usages. 66 p.

Amato G et Straccia U. User profile modeling and applications to digital libraries. Dans : International Conference on Theory and Practice of Digital Libraries, 1999a. Springer, p. 184-197.

Amato G et Straccia U. 1999b. User profile modeling and applications to digital libraries. Dans : Research and Advanced Technology for Digital Libraries. Springer, p. 184-197.

Angeletou S, Sabou M, Specia L et Motta E. 2007. Bridging the gap between folksonomies and the semantic web: An experience report.

Anil NK, Kurian SB et Varghese SM. 2013. Multidimensional User Data Model for Web Personalization. arXiv preprint arXiv:13064427.

Anuradha R. Kale PVTG, Prof. H.N. Datir. June-2013. Re-ranking the Results Based on user profile. International Journal of Advancements in Research & Technology, 2 : 5.

Aouicha MB. 2009. Une approche algébrique pour la recherche d'information structurée.

Asfari O. 2011. Personalized access to contextual information by using an assistant for query reformulation. Citeseer.

Atanassova I et Bertin M. 2014. Semantic facets for scientific information retrieval. Dans : Semantic Web Evaluation Challenge. Springer, p. 108-113.

Audeh B, Beaune P et Beigbeder M. La reformulation hybride des requêtes exploratoires à l'aide de concepts explicites et implicites. Dans : COnf'erence en Recherche d'Infomations et Applications-CORIA 2014, 11th French Information Retrieval Conference, 2014. p. pp. 247-260.

Auray N. 2007. Folksonomy: The new way to serendipity.

Baby B et Murali S. 2016. A SURVEY ON TRUST BASED RECOMMENDATION SYSTEMS.

Badache I. RI sociale: intégration de propriétés sociales dans un modèle de recherche. Dans : Conférence francophone en Recherche d'Information et Applications-CORIA 2013, 2013. p. pp. 1-6.

Baeza-Yates R et Ribeiro-Neto B. 1999. Modern information retrieval. ACM press New York.

Bao S, Xue G, Wu X, Yu Y, Fei B et Su Z. Optimizing web search using social annotations. Dans : Proceedings of the 16th international conference on World Wide Web, 2007. ACM, p. 501-510.

Barjasteh I, Forsati R, Masrour F, Esfahanian A-H et Radha H. Cold-start item and user recommendation with decoupled completion and transduction. Dans : Proceedings of the 9th ACM Conference on Recommender Systems, 2015. ACM, p. 91-98.

Barry CL. 1994. User-defined relevance criteria: an exploratory study. Journal of the American Society for Information Science, 45 : 149.

Baziz M. 2005. Indexation conceptuelle guidée par ontologie pour la recherche d'information. Toulouse 3.

Begg IM, Gnocato J et Moore WE. A prototype intelligent user interface for real-time supervisory control systems. Dans : Proceedings of the 1st international conference on Intelligent user interfaces, 1993. ACM, p. 211-214.

Beldjoudi S. 2015. La Sémantique et l'Effet Communautaire: Enrichissement et Exploitation. Université Pierre et Marie Curie Paris (France.

Beldjoudi S, Seridi H et Benzine A. Améliorer la Recommandation de Ressources dans les Folksonomies par l'Utilisation de Linked Open Data.

Beldjoudi S, Seridi H et Faron-Zucker C. Ambiguity in tagging and the community effect in researching relevant resources in folksonomies. Dans : Proc of ESWC Workshop User Profile Data on the Social Semantic Web, 2011a.

Beldjoudi S, Seridi H et Zucker CF. Improving tag-based resource recommendation with association rules on folksonomies. Dans : 2nd ISWC Workshop on Semantic Personalized Information Management: Retrieval and Recommendation, SPIM 2011, 2011b.

Beldjoudi S, Seridi-Bouchelaghem H et Faron-Zucker C. Personalizing and improving tag-based search in folksonomies. Dans : International Conference on Artificial Intelligence: Methodology, Systems, and Applications, 2012. Springer, p. 112-118.

Beldjoudi S, Seridi H et Benzine A. Améliorer la Recommandation de Ressources dans les Folksonomies par l'Utilisation de Linked Open Data. Dans : IC2016: Ingénierie des Connaissances, 2016.

Beldjoudi S, Seridi H et Faron Zucker C. 2017. PERSONALIZING AND IMPROVING RESOURCE RECOMMENDATION BY ANALYZING USERS PREFERENCES IN SOCIAL TAGGING ACTIVITIES. Computing & Informatics, 36.

Belkin NJ et Croft WB. 1992. Information filtering and information retrieval: Two sides of the same coin? Communications of the ACM, 35 : 29-38.

Bendakir N et Aïmeur E. Using association rules for course recommendation. Dans : Proceedings of the AAAI Workshop on Educational Data Mining, 2006.

Bender M, Crecelius T, Kacimi M, Michel S, Neumann T, Parreira JX, Schenkel R et Weikum G. Exploiting social relations for query expansion and result ranking. Dans : Data engineering workshop, 2008 ICDEW 2008 IEEE 24th International Conference on, 2008. IEEE, p. 501-506.

Benhamdi S, Babouri A et Chiky R. 2017. Personalized recommender system for e-Learning environment. Education and Information Technologies, 22 : 1455-1477.

Berrueta D, Labra JE et Polo L. 2006. Searching over public administration legal documents using ontologies. Frontiers in artificial intelligence and applications, 140 : 167.

Bertier M, Guerraoui R, Leroy V et Kermarrec A-M. Toward personalized query expansion. Dans : Proceedings of the Second ACM EuroSys Workshop on Social Network Systems, 2009. ACM, p. 7-12.

Bhogal J, MacFarlane A et Smith P. 2007. A review of ontology based query expansion. Information processing & management, 43 : 866-886.

Biancalana C et Micarelli A. Social tagging in query expansion: A new way for personalized web search. Dans : Computational Science and Engineering, 2009 CSE'09 International Conference on, 2009. IEEE, p. 1060-1065.

Biancalana C, Lapolla A et Micarelli A. Personalized web search using correlation matrix for query expansion. Dans : International Conference on Web Information Systems and Technologies, 2008. Springer, p. 186-198.

Bonnel N et Moreau F. Quel avenir pour les moteurs de recherche? Dans : MajecSTIC 2005: Manifestation des Jeunes Chercheurs francophones dans les domaines des STIC, 2005. p. 291-299.

Borlund P et Ingwersen P. Measures of relative relevance and ranked half-life: performance indicators for interactive IR. Dans : Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval, 1998. ACM, p. 324-331.

Borzsony S, Kossmann D et Stocker K. The skyline operator. Dans : Data Engineering, 2001 Proceedings 17th International Conference on, 2001. IEEE, p. 421-430.

Bouadjenek MR, Hacid H, Bouzeghoub M et Vakali A. Using social annotations to enhance document representation for personalized search. Dans : Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval, 2013. ACM, p. 1049-1052.

Bouadjenek MR, Hacid H, Bouzeghoub M et Vakali A. 2016. PerSaDoR: Personalized social document representation for improving web search. Information Sciences, 369 : 614-633.

Boubekeur F, Boughanem M, Tamine L et Daoud M. Using WordNet for Concept-based document indexing in information retrieval. Dans : Fourth International Conference on Semantic Processing (SEMAPRO), Florence, Italy, 2010.

Boudiba T-R et Ahmed-Ouamer R. 2017. Approche temporelle pour la génération personnalisée de profils folksonomiques. INFORSID, p. 263-273.

Boughanem M, Kraaij W et Nie J-Y. 2004. Modeles de langue pour la recherche d'information. Les systemes de recherche d'informations : 163-182.

Boughareb D et Farah N. Contextual modelling of the user browsing behaviour to identify the user's information need. Dans : Second International Conference on the Innovative Computing Technology (INTECH 2012), 2012.

Boughareb D et Farah N. 2013a. A Query Expansion Approach Using the Context of the Search. Dans : Ambient Intelligence-Software and Applications. Springer, p. 57-63.

Boughareb D et Farah N. 2013b. Identify the User's Information Need Using the Current Search Context. International Journal of Enterprise Information Systems (IJEIS), 9 : 28-42.

Bouhini C. 2014. Impact des réseaux sociaux sur le processus de recherche d'information. Ecole Nationale Supérieure des Mines de Saint-Etienne.

Bouhini C, Géry M et Largeron C. Modèle de Recherche d'Information Sociale Centré Utilisateur. Dans : Extraction et gestion des connaissances (EGC'2013), 2013a. Hermann, p. 275-286.

Bouhini C, Géry M et Largeron C. User-Centered Social Information Retrieval Model Exploiting Annotations and Social Relationships. Dans : Asia Information Retrieval Symposium, 2013b. Springer, p. 356-367.

Bouhini C, Géry M et Largeron C. Personalized information retrieval models integrating the user's profile. Dans : Research Challenges in Information Science (RCIS), 2016 IEEE Tenth International Conference on, 2016. IEEE, p. 1-9.

Bouidghaghen O et Tamine L. 2012. Spatio-Temporal Based Personalization for Mobile search, Engineering, and Intelligent Technologies. Next Generation Search Engines: Advanced Models for Information Retrieval. PA: IGI Global publishing.

Bouklit M et Lafourcade M. Propagation de signatures lexicales dans le graphe du web. Dans : Proc of RFIA, 2006.

Bradley K, Rafter R et Smyth B. Case-based user profiling for content personalisation. Dans : International Conference on Adaptive Hypermedia and Adaptive Web-Based Systems, 2000. Springer, p. 62-72.

Brin S et Page L. 1998. The anatomy of a large-scale hypertextual web search engine. Computer networks and ISDN systems, 30 : 107-117.

Broder A. A taxonomy of web search. Dans : ACM Sigir forum, 2002. ACM, p. 3-10.

Buffa M, Gandon F, Ereteo G, Sander P et Faron C. 2008. SweetWiki: A semantic wiki. Web Semantics: Science, Services and Agents on the World Wide Web, 6 : 84-97.

Buyukokkten O, Cho J, Garcia-Molina H, Gravano L et Shivakumar N. 1999. Exploiting geographical location information of web pages.

Cai Y et Li Q. Personalized search by tag-based user profile and resource profile in collaborative tagging systems. Dans : Proceedings of the 19th ACM international conference on Information and knowledge management, 2010. ACM, p. 969-978.

Cai Y, Leung H-f, Li Q, Min H, Tang J et Li J. 2014. Typicality-based collaborative filtering recommendation. IEEE Transactions on knowledge and data engineering, 26 : 766-779.

Cailliau F. 2010. Des ressources aux traitements linguistiques: le rôle d ‘une architecture linguistique. Université Paris-Nord-Paris XIII.

Cakir O et Aras ME. 2012. A recommendation engine by using association rules. Procedia-Social and Behavioral Sciences, 62 : 452-456.

Callahan PB et Kosaraju SR. Faster Algorithms for Some Geometric Graph Problems in Higher Dimensions. Dans : SODA, 1993. p. 291-300.

Cantador I, Szomszor M, Alani H, Fernández M et Castells P. 2008. Enriching ontological user profiles with tagging history for multi-domain recommendations.

Carmagnola F, Cena F, Cortassa O, Gena C et Torre I. Towards a tag-based user model: How can user model benefit from tags? Dans : International Conference on User Modeling, 2007. Springer, p. 445-449.

Carmel D, Uziel E, Guy I, Mass Y et Roitman H. 2012. Folksonomy-based term extraction for word cloud generation. ACM Transactions on Intelligent Systems and Technology (TIST), 3 : 60.

Carmel D, Zwerdling N, Guy I, Ofek-Koifman S, Har'El N, Ronen I, Uziel E, Yogev S et Chernov S. Personalized social search based on the user's social network. Dans : Proceedings of the 18th ACM conference on Information and knowledge management, 2009. ACM, p. 1227-1236.

Carpineto C et Romano G. 2012. A survey of automatic query expansion in information retrieval. ACM Computing Surveys (CSUR), 44 : 1.

Challam V, Gauch S et Chandramouli A. Contextual search using ontology-based user profiles. Dans : Large Scale Semantic Access to Content (Text, Image, Video, and Sound), 2007. LE CENTRE DE HAUTES ETUDES INTERNATIONALES D'INFORMATIQUE DOCUMENTAIRE, p. 612-617.

Chandrakar O et Saini JR. 2015. Predicting Examination Results using Association Rule Mining. International Journal of Computer Applications, 116.

Chekkai N, Chikhi S et Kheddouci H. Nouvelle Approche à base de Graphes pour les Systèmes de Recommandation Collaboratifs. Dans : CIIA, 2011. Citeseer.

Chekkai N, Chikhi S et Kheddouci H. A weighted-graph based approach for solving the cold start problem in collaborative recommender systems. Dans : Computers and Communications (ISCC), 2012 IEEE Symposium on, 2012. IEEE, p. 000759-000764.

Chekkai N, Chikhi S et Kheddouci H. 2013. Weighted graph-based methods for identifying the most influential actors in trust social networks. International Journal of Networking and Virtual Organisations, 13 : 101-128.

Chen CC, Wan Y-H, Chung M-C et Sun Y-C. 2013. An effective recommendation method for cold start new users using trust and distrust networks. Information Sciences, 224 : 19-36.

Cheng C, Angustia T, Ching MH, Cristobal CA et Gabuyo GM. 2014. Synonym Based Tag Cloud Generation.

Cherniack M, Galvez EF, Franklin MJ et Zdonik S. Profile-driven cache management. Dans : Data Engineering, 2003 Proceedings 19th International Conference on, 2003. IEEE, p. 645-656.

Chirita P-A, Firan CS et Nejdl W. Personalized query expansion for the web. Dans : Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval, 2007. ACM, p. 7-14.

Chirita PA, Nejdl W, Paiu R et Kohlschütter C. Using ODP metadata to personalize search. Dans : Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval, 2005. ACM, p. 178-185.

Choi S-S, Cha S-H et Tappert CC. 2010. A survey of binary similarity and distance measures. Journal of Systemics, Cybernetics and Informatics, 8 : 43-48.

Clauset A. 2005. Finding local community structure in networks. Physical review E, 72 : 026132.

Cleverdon C. The Cranfield tests on index language devices. Dans : Aslib proceedings, 1967. MCB UP Ltd, p. 173-194.

Cleverdon C. 1970. Evaluation tests of information retrieval systems. Journal of Documentation, 26 : 55-67.

Condli MK, Lewis DD, Madigan D et Posse C. Bayesian Mixed-E ects Models for Recommender Systems. Dans : ACM SIGIR, 1999.

Corby O, Dieng-Kuntz R et Faron-Zucker C. Querying the semantic web with corese search engine. Dans : ECAI, 2004. p. 705.

Corby O, Dieng-Kuntz R, Gandon F et Faron-Zucker C. 2006. Searching the semantic web: Approximate query processing based on ontologies. Intelligent Systems, IEEE, 21 : 20-27.

Cuadra CA et Katter RV. 1967. Opening the black box of 'relevance'. Journal of Documentation, 23 : 291-303.

Cuong BC et Long HV. 2013. Spatial interaction–modification model and applications to geo-demographic analysis. Knowledge-Based Systems, 49 : 152-170.

Cuong BC, Lanzi PL et Thong NT. 2012. A novel intuitionistic fuzzy clustering method for geo-demographic analysis. Expert Systems with Applications, 39 : 9848-9859.

da Silva STF, de Oliveira Apolonio S, Vivacqua AS, Oliveira J, Xexéo GB et Campos MLM. Ontoogle: Enhancing retrieval with ontologies and facets. Dans : Computer Supported Cooperative Work in Design (CSCWD), 2011 15th International Conference on, 2011. IEEE, p. 192-199.

DAFT R et HUBER O. 1975. A review of an a framework for the thinking on the notion in Information Science. Journal of the American Society for Information Science : 321-343.

Daoud M. 2009. Accès personnalisé à l'information: approche basée sur l'utilisation d'un profil utilisateur sémantique dérivé d'une ontologie de domaines à travers l'historique des sessions de recherche. Université Paul Sabatier-Toulouse III.

Daoud M, Tamine-Lechani L et Boughanem M. Learning user interests for a session-based personalized search. Dans : Proceedings of the second international symposium on Information interaction in context, 2008. ACM, p. 57-64.

Daoud M, Tamine L et Chebaro B. 2010a. Proposition d'un système de RI personnalisé à base de sessions intégrant un profil utilisateur sémantique.

Daoud M, Tamine L et Chebaro B. 2010b. Proposition d'un système de RI personnalisé à base de sessions intégrant un profil utilisateur sémantique. Document numérique, 13 : 137-160.

Daoud M, Tamine-Lechani L, Boughanem M et Chebaro B. A session based personalized search using an ontological user profile. Dans : Proceedings of the 2009 ACM symposium on Applied Computing, 2009. ACM, p. 1732-1736.

De Meo P, Quattrone G et Ursino D. 2010. A query expansion and user profile enrichment approach to improve the performance of recommender systems operating on a folksonomy. User Modeling and User-Adapted Interaction, 20 : 41-86.

Deerwester S, Dumais ST, Furnas GW, Landauer TK et Harshman R. 1990a. Indexing by latent semantic analysis. Journal of the American Society for Information Science, 41 : 391.

Deerwester SC, Dumais ST, Landauer TK, Furnas GW et Harshman RA. 1990b. Indexing by latent semantic analysis. JASIS, 41 : 391-407.

Desmontils E et Jacquin C. Indexing a web site with a terminology oriented ontology. Dans : Proceedings of the First International Conference on Semantic Web Working, 2001. CEUR-WS. org, p. 549-565.

Desmontils E, Jacquin C et Morin E. 2002. Indexation sémantique de documents sur le Web: application aux ressources humaines. Proceedings of Journées de l'AS-CNRS Web sémantique.

Desrosiers C et Karypis G. 2011. A comprehensive survey of neighborhood-based recommendation methods. Dans : Recommender systems handbook. Springer, p. 107-144.

Di Noia T et Ostuni VC. Recommender systems and linked open data. Dans : Reasoning Web International Summer School, 2015. Springer, p. 88-113.

Diaz F. Integration of news content into web results. Dans : Proceedings of the Second ACM International Conference on Web Search and Data Mining, 2009. ACM, p. 182-191.

Dillon M. 1983. Introduction to modern information retrieval: G. Salton and M. McGill. McGraw-Hill, New York (1983). xv+ 448 pp., $32.95 ISBN 0-07-054484-0. Pergamon.

Dinh D et Tamine L. 2012. Towards a context sensitive approach to searching information based on domain specific knowledge sources. Web Semantics: Science, Services and Agents on the World Wide Web, 12 : 41-52.

Djalila B. 2014. Recherche d'information multicritères. Université Badji Mokhtar de Annaba.

Domneti R. 2009. Neighborhood based methods for Collaborative Filtering. A Case Study, I : 1-5.

Du Q, Xie H, Cai Y, Leung H-f, Li Q, Min H et Wang FL. 2016. Folksonomy-based personalized search by hybrid user profiles in multiple levels. Neurocomputing, 204 : 142-152.

Dugast C. 2011. Publiée une fois par année, la Revue électronique suisse de science de l'information (RESSI) a pour but principal le développement scientifique de cette discipline en Suisse. Ressi.

Dumais S. Evaluating IR in situ. Dans : Proceedings of the SIGIR 2009 Workshop on the Future of IR Evaluation, 2009. p. 2.

Durao F et Dolog P. 2012. A personalized tag-based recommendation in social web systems. arXiv preprint arXiv:12030332.

Ehrig M, Haase P, Hefke M et Stojanovic N. 2005. Similarity for ontologies-a comprehensive framework. ECIS 2005 Proceedings : 127.

English J, Hearst M, Sinha R, Swearingen K et Lee K. 2002. Flexible search and navigation using faceted metadata. Technical report, University of Berkeley, School of Information Management and Systems, 2003. Submitted for publication.

Evéquoz F, Thomet J et Lalanne D. Gérer son information personnelle au moyen de la navigation par facettes. Dans : Conference Internationale Francophone sur l'Interaction Homme-Machine, 2010. ACM, p. 41-48.

Fagan JC. 2013. Usability studies of faceted browsing: A literature review. Information Technology and Libraries : 58.

Ferber R. Using Co-occurrence Data for Query Expansion: Wrong Paradigm or Wrong Formulas.

Fernández M-L et Valiente G. 2001. A graph distance metric combining maximum common subgraph and minimum common supergraph. Pattern Recognition Letters, 22 : 753-758.

Ferschke O, Zesch T et Gurevych I. Wikipedia revision toolkit: efficiently accessing Wikipedia's edit history. Dans : Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Systems Demonstrations, 2011. Association for Computational Linguistics, p. 97-102.

Fonseca BM, Golgher P, Pôssas B, Ribeiro-Neto B et Ziviani N. Concept-based interactive query expansion. Dans : Proceedings of the 14th ACM international conference on Information and knowledge management, 2005. ACM, p. 696-703.

Fox EA. 1983. Extending the boolean and vector space models of information retrieval with p-norm queries and multiple concept types.

Fu KAH et Kim H. Personalizing search: A case for scaling concurrency in multitenant semantic web search systems. Dans : IEEE International Conference on Big Data, 2013.

Gauch S, Chaffee J et Pretschner A. 2003a. Ontology-based personalized search and browsing. Web Intelligence and Agent Systems: An international Journal, 1 : 219-234.

Gauch S, Speretta M, Chandramouli A et Micarelli A. 2007. User profiles for personalized information access. Dans : The adaptive web. Springer, p. 54-89.

Gauch S, Madrid JM, Induri S, Ravindran D et Chadlavada S. 2003b. Keyconcept: A conceptual search engine. Information and Telecommunication Technology Center.

Geetharani S et Soranamageswari M. Location-based Ranking Method (LBRM) for ranking search results in search engines. Dans : Intelligent Systems and Control (ISCO), 2016 10th International Conference on, 2016. IEEE, p. 1-6.

Gemmell J, Shepitsen A, Mobasher B et Burke R. 2008. Personalization in folksonomies based on tag clustering. Intelligent techniques for web personalization & recommender systems, 12.

Géry M, Largeron C et Thollard F. 2010. BM25t, une extension de BM25 pour la recherche d'information ciblée. Document numérique, 13 : 83-110.

Gibson D, Kleinberg J et Raghavan P. Inferring web communities from link topology. Dans : Proceedings of the ninth ACM conference on Hypertext and hypermedia: links, objects, time and space---structure in hypermedia systems: links, objects, time and space---structure in hypermedia systems, 1998. ACM, p. 225-234.

Gödert W. 2014. Facets and Typed Relations as Tools for Reasoning Processes in Information Retrieval. Dans : Metadata and Semantics Research. Springer, p. 128-140.

Godin R, Missaoui R et April A. 1993. Experimental comparison of navigation in a Galois lattice with conventional information retrieval methods. International Journal of Man-Machine Studies, 38 : 747-767.

Godoy D et Amandi A. Hybrid content and tag-based profiles for recommendation in collaborative tagging systems. Dans : Latin American Web Conference, 2008 LA-WEB'08, 2008. IEEE, p. 58-65.

Golbeck J et Hendler J. Filmtrust: Movie recommendations using trust in web-based social networks. Dans : Proceedings of the IEEE Consumer communications and networking conference, 2006. p. 282-286.

Gong Z et Cheang CW. Multi-term web query expansion using WordNet. Dans : International Conference on Database and Expert Systems Applications, 2006. Springer, p. 379-388.

Gonzalez G, De La Rosa JL, Montaner M et Delfin S. Embedding emotional context in recommender systems. Dans : Data Engineering Workshop, 2007 IEEE 23rd International Conference on, 2007. IEEE, p. 845-852.

Gonzalo J, Verdejo F, Chugur I et Cigarran J. 1998. Indexing with WordNet synsets can improve text retrieval. arXiv preprint cmp-lg/9808002.

Gormley C et Tong Z. 2015. Elasticsearch: The Definitive Guide. " O'Reilly Media, Inc.".

Granovetter MS. 1973. The strength of weak ties. American journal of sociology, 78 : 1360-1380.

Group CR et Vickery BC. 1963. La classification à facettes: guide pour la construction et l'utilisation de schémas spéciaux, rédigé. Gauthier-Villars, 1963 [ie 1962].

Guarino N, Masolo C et Vetere G. 1999. Ontoseek: Content-based access to the web. IEEE Intelligent Systems and their Applications, 14 : 70-80.

Guo G. Integrating trust and similarity to ameliorate the data sparsity and cold start for recommender systems. Dans : Proceedings of the 7th ACM conference on Recommender systems, 2013. ACM, p. 451-454.

Hannech A, Adda M et Mcheick H. Multi-space Projection Based Search Engine: Theoretical Model Instantiation and Prototype. Dans : Database and Expert Systems Applications (DEXA), 2015 26th International Workshop on, 2015. IEEE, p. 281-285.

Hannech A, Adda M et Mcheick H. Recommendation Model Based on a Contextual Similarity Measure. Dans : Machine Learning and Applications (ICMLA), 2016 15th IEEE International Conference on, 2016a. IEEE, p. 394-401.

Hannech A, Adda M et Mcheick H. Cold-start recommendation strategy based on social graphs. Dans : Information Technology, Electronics and Mobile Communication Conference (IEMCON), 2016 IEEE 7th Annual, 2016b. IEEE, p. 1-7.

Hannech A, Adda M et Mcheick H. Social data-based user profile enrichment. Dans : Information Technology, Electronics and Mobile Communication Conference (IEMCON), 2016 IEEE 7th Annual, 2016c. IEEE, p. 1-7.

Harter SP. 1992. Psychological relevance and information science. Journal of the American Society for Information Science (1986-1998), 43 : 602.

Harter SP et Hert CA. 1997. Evaluation of information retrieval systems: Approaches, issues, and methods. Annual Review of Information Science and Technology (ARIST), 32 : 3-94.

Hascoët M et Beaudouin-Lafon M. 2001. Visualisation interactive d'information. Revue I3, 1 : 77-108.

Hassan-Montero Y et Herrero-Solana V. Improving tag-clouds as visual information retrieval interfaces. Dans : International conference on multidisciplinary information sciences and technologies, 2006. p. 25-28.

Haveliwala TH, Gionis A, Klein D et Indyk P. Evaluating strategies for similarity search on the web. Dans : Proceedings of the 11th international conference on World Wide Web, 2002. ACM, p. 432-442.

Hawalah A et Fasli M. 2015. Dynamic user profiles for web personalisation. Expert Systems with Applications, 42 : 2547-2569.

Haydar C, Boyer A et Roussanaly A. Hybridising collaborative filtering and trust-aware recommender systems. Dans : 8th International Conference on Web Information Systems and Technologies-WEBIST'2012, 2012.

Hearst M. Design recommendations for hierarchical faceted search interfaces. Dans : ACM SIGIR workshop on faceted search, 2006. Seattle, WA, p. 1-5.

Hildebrand M, van Ossenbruggen J et Hardman L. 2006. /facet: A browser for heterogeneous semantic web repositories. Springer.

Hirst G et St-Onge D. 1998. Lexical chains as representations of context for the detection and correction of malapropisms. WordNet: An electronic lexical database, 305 : 305-332.

Hmimida M. 2012. Approche de recommandation de tags par niveau. 3ième conférence sur les modèles et l'analyse des réseaux : Approches mathématiques et informatiques.

Hmimida M et Kanawati R. A Graph-Coarsening Approach for Tag Recommendation. Dans : Proceedings of the 25th International Conference Companion on World Wide Web, 2016. International World Wide Web Conferences Steering Committee, p. 43-44.

Hotho A, Jäschke R, Schmitz C et Stumme G. Information retrieval in folksonomies: Search and ranking. Dans : European Semantic Web conference, 2006. Springer, p. 411-426.

Hsu I-C. 2013. Integrating ontology technology with folksonomies for personalized social tag recommendation. Applied Soft Computing, 13 : 3745-3750.

Hu Y, Xin G, Song R, Hu G, Shi S, Cao Y et Li H. extraction from bodies of HTML documents and its application to web page retrieval. Dans : Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval, 2005. ACM, p. 250-257.

Huang C-L, Chien H-Y et Conyette M. Folksonomy-based Recommender Systems with User-s Recent Preferences. Dans : International Conference on Computer and Information Science and Engineering, Amsterdam, Netherlands, 2011.

Huang C-L, Yeh P-H, Lin C-W et Wu D-C. 2014. Utilizing user tag-based interests in recommender systems for social resource sharing websites. Knowledge-Based Systems, 56 : 86-96.

Huang EH, Socher R, Manning CD et Ng AY. Improving word representations via global context and multiple word prototypes. Dans : Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1, 2012. Association for Computational Linguistics, p. 873-882.

HUDON M. 1997. INDEXATION ET LANGAGES DOCUMENTAIRES DANS LES MILIEUX ARCBIVISTIQUES A L'ERE DES NOUVELLES TECHNOLOGIES DE L'INFORMATION. Archives, 29 : 75-98.

Ihadjadene M. 2004. Les systèmes de recherche d'informations: modèles conceptuels. Hermès Science.

Imafouo A et Tannier X. Retrieval status values in information retrieval evaluation. Dans : International Symposium on String Processing and Information Retrieval, 2005. Springer, p. 224-227.

Jabeur LB, Tamine L et Boughanem M. Un modèle de Recherche d'Information Sociale pour l'Accès aux Ressources Bibliographiques: Vers un réseau social pondéré. Dans : Atelier REcherche et REcommandation d'information dans les RESeaux sOciaux à INFORSID 2010, 2010. p. 37-49.

Jain A, Mittal K et Sabharwal S. Conceptual Weighing Query Expansion based on User Profiles.

Jamali M et Ester M. Trustwalker: a random walk model for combining trust-based and item-based recommendation. Dans : Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining, 2009. ACM, p. 397-406.

Jannach D, Zanker M, Felfernig A et Friedrich G. 2010. Recommender systems: an introduction. Cambridge University Press.

Jansen BJ et Spink A. 2006. How are we searching the World Wide Web? A comparison of nine search engine transaction logs. Information processing & management, 42 : 248-263.

Jansen BJ, Spink A et Saracevic T. 2000. Real life, real users, and real needs: a study and analysis of user queries on the web. Information processing & management, 36 : 207-227.

Jaro MA. 1989. Advances in record-linkage methodology as applied to matching the 1985 census of Tampa, Florida. Journal of the American Statistical Association, 84 : 414-420.

Jäschke R, Marinho L, Hotho A, Schmidt-Thieme L et Stumme G. Tag recommendations in folksonomies. Dans : European Conference on Principles of Data Mining and Knowledge Discovery, 2007. Springer, p. 506-514.

Jaseena K et David JM. 2014. Issues, challenges, and solutions: big data mining. NeTCoM, CSIT, GRAPH-HOC, SPTM–2014 : 131-140.

Jelassi MN, Yahia SB et Nguifo EM. Vers des recommandations plus personnalisées dans les folksonomies. Dans : IC-25èmes Journées francophones d'Ingénierie des Connaissances, 2014. p. 187-198.

Jelassi MN, Yahia SB et Nguifo EM. PersoRec: un système personnalisé de recommandations pour les folksonomies basé sur les concepts quadratiques. Dans : EGC, 2016. p. 487-492.

Jiang M, Cui P, Liu R, Yang Q, Wang F, Zhu W et Yang S. Social contextual recommendation. Dans : Proceedings of the 21st ACM international conference on Information and knowledge management, 2012. ACM, p. 45-54.

Joly A, Maret P et Daigremont J. Contextual recommendation of social updates, a tag-based framework. Dans : International Conference on Active Media Technology, 2010. Springer, p. 436-447.

Jomsri P, Sanguansintukul S et Choochaiwattana W. Improving research paper searching with social tagging—A preliminary investigation. Dans : Natural Language Processing, 2009 SNLP'09 Eighth International Symposium on, 2009. IEEE, p. 152-156.

Ju C et Xu C. 2013. A new collaborative recommendation approach based on users clustering using artificial bee colony algorithm. The Scientific World Journal, 2013.

Kang I-H et Kim G. Query type classification for web document retrieval. Dans : Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval, 2003. ACM, p. 64-71.

Karbasi S. 2007. Pondération des termes en recherche d'information: modèle de pondération basé sur le rang des termes dans les documents. Insitut de recherche en informatique de Toulouse.

Kayser D. 1997. La représentation des connaissances. Hermes.

Kessler MM. 1963. Bibliographic coupling between scientific papers. Journal of the Association for Information Science and Technology, 14 : 10-25.

Kessler MM. 1965. Comparison of the results of bibliographic coupling and analytic subject indexing. Journal of the Association for Information Science and Technology, 16 : 223-233.

Khalefa ME, Mokbel MF et Levandoski JJ. Skyline query processing for incomplete data. Dans : Data Engineering, 2008 ICDE 2008 IEEE 24th International Conference on, 2008. IEEE, p. 556-565.

Khelif K et Dieng-Kuntz R. Annotations sémantiques pour le domaine Biopuces. Dans : 15èmes Journées francophones d'Ingénierie des Connaissances, 2004. Presses universitaires de Grenoble, p. 273-284.

Kießling W. Foundations of preferences in database systems. Dans : Proceedings of the 28th international conference on Very Large Data Bases, 2002. VLDB Endowment, p. 311-322.

Kim H-N, Alkhaldi A, El Saddik A et Jo G-S. 2011. Collaborative user modeling with user-generated tags for social recommender systems. Expert Systems with Applications, 38 : 8488-8496.

Kim HR et Chan PK. Learning implicit user interest hierarchy for context in personalization. Dans : Proceedings of the 8th international conference on Intelligent user interfaces, 2003. ACM, p. 101-108.

Kleinberg JM. 1999. Authoritative sources in a hyperlinked environment. Journal of the ACM (JACM), 46 : 604-632.

Koll MB. WEIRD: An approach to concept-based information retrieval. Dans : ACM Sigir Forum, 1979. ACM, p. 32-50.

Konstas I, Stathopoulos V et Jose JM. On social networks and collaborative recommendation. Dans : Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval, 2009. ACM, p. 195-202.

Koren J, Zhang Y et Liu X. Personalized interactive faceted search. Dans : Proceedings of the 17th international conference on World Wide Web, 2008. ACM, p. 477-486.

Kostadinov D. 2007. Data Personalization: an approach for profile management and query reformulation. PhD thesis, University of Versailles, France.

Koutrika G et Ioannidis Y. A unified user profile framework for query disambiguation and personalization. Dans : Proceedings of workshop on new technologies for personalized information access, 2005. p. 44-53.

Kraft DH et Buell DA. 1983. Fuzzy sets and generalized Boolean retrieval systems. International Journal of Man-Machine Studies, 19 : 45-56.

Krovetz R. Homonymy and polysemy in information retrieval. Dans : Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics, 1997. Association for Computational Linguistics, p. 72-79.

Krovetz R et Croft WB. Word sense disambiguation using machine-readable dictionaries. Dans : ACM SIGIR Forum, 1989. ACM, p. 127-136.

Kumar N et Carterette B. Time based feedback and query expansion for twitter search. Dans : European Conference on Information Retrieval, 2013. Springer, p. 734-737.

Kumar R, Raghavan P, Rajagopalan S et Tomkins A. 1999. Trawling the Web for emerging cyber-communities. Computer networks, 31 : 1481-1493.

Kwok K. A neural network for probabilistic information retrieval. Dans : ACM SIGIR Forum, 1989. ACM, p. 21-30.

Kwok K. 1995. A network approach to probabilistic information retrieval. ACM Transactions on Information Systems (TOIS), 13 : 324-353.

Labrou Y et Finin T. Yahoo! as an ontology: using Yahoo! categories to describe documents. Dans : Proceedings of the eighth international conference on Information and knowledge management, 1999. ACM, p. 180-187.

Lacoste C, Chevallet J-P, Lim J-H, Wei X, Raccoceanu D, Hoang DLT, Teodorescu R et Vuillenemot N. Ipal knowledge-based medical image retrieval in imageclefmed 2006. Dans : Working Notes for the CLEF 2006 Workshop, 2006. Citeseer, p. 20-22.

Lanzi PL, Cuong BC et Hung HA. 2012. Data Mining in GIS: A Novel Context-Based Fuzzy Geographically Weighted ClusteringAlgorithm. International Journal of Machine Learning and Computing, 2 : 235.

Lavrenko V et Croft WB. Relevance based language models. Dans : Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval, 2001. ACM, p. 120-127.

Le T, Vo B et Duong TH. Personalized facets for semantic search using linked open data with social networks. Dans : Innovations in Bio-Inspired Computing and Applications (IBICA), 2012 Third International Conference on, 2012. IEEE, p. 312-317.

Lee DH et Brusilovsky P. Social networks and interest similarity: the case of CiteULike. Dans : Proceedings of the 21st ACM conference on Hypertext and hypermedia, 2010. ACM, p. 151-156.

Lee J, Sun M et Lebanon G. 2012. A comparative study of collaborative filtering algorithms. arXiv preprint arXiv:12053193.

Lee U, Liu Z et Cho J. Automatic identification of user goals in web search. Dans : Proceedings of the 14th international conference on World Wide Web, 2005. ACM, p. 391-400.

Li J, Tang B et Cercone N. Applying association rules for interesting recommendations using rule templates. Dans : Pacific-Asia Conference on Knowledge Discovery and Data Mining, 2004. Springer, p. 166-170.

Li L, Peng W, Kataria S, Sun T et Li T. 2015. Recommending users and communities in social media. ACM Transactions on Knowledge Discovery from Data (TKDD), 10 : 17.

Li Y et Zhong N. Ontology based web mining for information gathering. Dans : International Workshop on Web Intelligence Meets Brain Informatics, 2006. Springer, p. 406-427.

Li Y et Belkin NJ. 2008. A faceted approach to conceptualizing tasks in information seeking. Information processing & management, 44 : 1822-1837.

Limpens F, Gandon F et Buffa M. Rapprocher les ontologies et les folksonomies pour la gestion des connaissances partagées: un état de l'art. Dans : 19es Journées Francophones d'Ingénierie des Connaissances (IC 2008), 2008. p. 123-134.

Lin C, Xue G-R, Zeng H-J et Yu Y. 2005. Using probabilistic latent semantic analysis for personalized web search. Web Technologies Research and Development-APWeb 2005 : 707-717.

Lin D. An information-theoretic definition of similarity. Dans : ICML, 1998. Citeseer, p. 296-304.

Lin S-H, Shih C-S, Chen MC, Ho J-M, Ko M-T et Huang Y-M. Extracting classification knowledge of Internet documents with mining term associations: a semantic approach. Dans : Proceedings of the 21st

annual international ACM SIGIR conference on Research and development in information retrieval, 1998. ACM, p. 241-249.

Lin W. 2000. Association rule mining for collaborative recommender systems. Citeseer.

Lin W, Alvarez SA et Ruiz C. 2002. Efficient adaptive-support association rule mining for recommender systems. Data mining and knowledge discovery, 6 : 83-105.

Lin Y, Lin H, Jin S et Ye Z. Social annotation in query expansion: a machine learning approach. Dans : Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval, 2011. ACM, p. 405-414.

Lipczak M. 2008. Tag recommendation for folksonomies oriented towards individual users. ECML PKDD discovery challenge, 84 : 2008.

Liu F, Yu C et Meng W. 2004a. Personalized web search for improving retrieval effectiveness. IEEE Transactions on knowledge and data engineering, 16 : 28-40.

Liu F, Yu C et Meng W. 2004b. Personalized web search for improving retrieval effectiveness. Knowledge and Data Engineering, IEEE transactions on, 16 : 28-40.

Liu Y, Zhang M, Ru L et Ma S. Automatic query type identification based on click through information. Dans : Asia Information Retrieval Symposium, 2006. Springer, p. 593-600.

Liu Y, Li C, Zhang P et Xiong Z. A query expansion algorithm based on phrases semantic similarity. Dans : Information Processing (ISIP), 2008 International Symposiums on, 2008. IEEE, p. 31-35.

Lops P, De Gemmis M et Semeraro G. 2011. Content-based recommender systems: State of the art and trends. Dans : Recommender systems handbook. Springer, p. 73-105.

Lops P, De Gemmis M, Semeraro G, Musto C et Narducci F. 2013. Content-based and collaborative techniques for tag recommendation: an empirical evaluation. Journal of Intelligent Information Systems, 40 : 41-61.

Lu C, Lam W et Zhang Y. Twitter user modeling and tweets recommendation based on wikipedia concept graph. Dans : Workshops at the Twenty-Sixth AAAI Conference on Artificial Intelligence, 2012.

Lu W, Robertson S et MacFarlane A. 2006. Field-weighted XML retrieval based on BM25. Advances in XML Information Retrieval and Evaluation : 161-171.

Lu Y-T, Yu S-I, Chang T-C et Hsu JY-j. A Content-Based Method to Enhance Tag Recommendation. Dans : IJCAI, 2009. p. 2064-2069.

Luo C, Liu Y, Zhang M et Ma S. Query ambiguity identification based on user behavior information. Dans : Asia Information Retrieval Symposium, 2014. Springer, p. 36-47.

Lv Y et Zhai C. Positional relevance model for pseudo-relevance feedback. Dans : Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval, 2010. ACM, p. 579-586.

Ma Y, Zeng Y, Ren X et Zhong N. User interests modeling based on multi-source personal information fusion and semantic reasoning. Dans : International Conference on Active Media Technology, 2011. Springer, p. 195-205.

Maedche A et Staab S. 2001. Ontology learning for the semantic web. IEEE Intelligent systems, 16 : 72-79.

Maguitman AG, Menczer F, Roinestad H et Vespignani A. Algorithmic detection of semantic similarity. Dans : Proceedings of the 14th international conference on World Wide Web, 2005. ACM, p. 107-116.

Maisonnasse L, Gaussier E et Chevallet JP. 2008. Multiplying concept sources for graph modeling. Dans : Advances in Multilingual and Multimodal Information Retrieval. Springer, p. 585-592.

Maisonnasse L, Gaussier E et Chevallet J-P. Combinaison d'analyses sémantiques pour la recherche d'information médicale. Dans : RISE (Recherche d'Information SEmantique) dans le cadre de la conférence INFORSID'2009, 2009.

Maloof MA et Michalski RS. 2000. Selecting examples for partial memory learning. Machine Learning, 41 : 27-52.

Manvitha V et Reddy MS. 2014. Music Recommendation System Using Association Rule Mining and Clustering Technique To Address Coldstart Problem. IJECS, 3 : 6855-6858.

Manzat A-M, Grigoras R et Sèdes F. Towards a user-aware enrichment of multimedia metadata. Dans : Workshop on Semantic Multimedia Database Technologies, 2010.

Marchiori M. 1998. The limits of Web metadata, and beyond. Computer networks and ISDN systems, 30 : 1-9.

Marleau Y, Mas S et Zacklad M. 2008. Exploitation des facettes et des ontologies sémiotiques pour la gestion documentaire. Traitements et pratiques documentaires: vers un changement de paradigme : 91-110.

Marsh SP. 1994. Formalising trust as a computational concept.

Mc Gowan JP. 2003. A multiple model approach to personalised information access. Citeseer.

Meng C, Cheng Y, Jiechao C et Peng Y. A Method to Solve Cold-Start Problem in Recommendation System based on Social Network Sub-community and Ontology Decision Model. Dans : 3rd International Conference on Multimedia Technology (ICMT-13), 2013. Atlantis Press.

Meo Pd, Ferrara E, Abel F, Aroyo L et Houben G-J. 2013. Analyzing user behavior across social sharing environments. ACM Transactions on Intelligent Systems and Technology (TIST), 5 : 14.

Merriam. 2008. Compounds words.

Mezghani M, Zayani CA, Amous I, Péninou A et Sedes F. Dynamic enrichment of social users' interests. Dans : Research Challenges in Information Science (RCIS), 2014 IEEE Eighth International Conference on, 2014. IEEE, p. 1-11.

Mican D et Tomai N. Association-rules-based recommender system for personalization in adaptive web-based applications. Dans : International Conference on Web Engineering, 2010. Springer, p. 85-90.

Micarelli A et Sciarrone F. 2004. Anatomy and empirical evaluation of an adaptive web-based information filtering system. User Modeling and User-Adapted Interaction, 14 : 159-200.

Michlmayr E et Cayzer S. 2007. Learning user profiles from tagging data and leveraging them for personal (ized) information access.

Michlmayr E, Cayzer S et Shabajee P. 2007. Add-A-Tag: Learning adaptive user profiles from bookmark collections.

Mihalkova L et Mooney R. Learning to disambiguate search queries from short sessions. Dans : Joint European Conference on Machine Learning and Knowledge Discovery in Databases, 2009. Springer, p. 111-127.

Mika P. Ontologies are us: A unified model of social networks and semantics. Dans : International semantic web conference, 2005. Springer, p. 522-536.

Mishne G. Autotag: a collaborative approach to automated tag assignment for weblog posts. Dans : Proceedings of the 15th international conference on World Wide Web, 2006. ACM, p. 953-954.

Missaoui R, Valtchev P, Djeraba C et Adda M. 2007. Toward recommendation based on ontology-powered web-usage mining. IEEE Internet Computing, 11.

Mizzaro S. 1997. Relevance: The whole history. JASIS, 48 : 810-832.

Mohamed S et Abdelmoty A. Uncovering User Profiles in Location-Based Social Networks. Dans : GEOProcessing 2016: The Eighth International Conference on Advanced Geographic Information Systems, Applications, and Services, 2016. p. 14-21.

Mostafa J, Mukhopadhyay S et Palakal M. 2003. Simulation studies of different dimensions of users' interests and their impact on user modeling and information filtering. Information retrieval, 6 : 199-223.

Musiał K et Kazienko P. 2013. Social networks on the internet. World Wide Web, 16 : 31-72.

Musto C, Narducci F, Lops P, De Gemmis M et Semeraro G. Content-based personalization services integrating folksonomies. Dans : International Conference on Electronic Commerce and Web Technologies, 2009a. Springer, p. 217-228.

Musto C, Narducci F, De Gemmis M, Lops P et Semeraro G. 2009b. A tag recommender system exploiting user and community behavior. Recommender Systems & the Social Web.

Nasir Uddin M et Janecek P. 2007. Performance and usability testing of multidimensional taxonomy in web site search and navigation. Performance measurement and metrics, 8 : 18-33.

Navigli R. 2009. Word sense disambiguation: A survey. ACM Computing Surveys (CSUR), 41 : 10.

Nguyen A-T, Denos N et Berrut C. Exploitation des données" disponibles à froid" pour améliorer le démarrage à froid dans les systèmes de filtrage d'information. Dans : INFORSID'06, 2006. p. 81--95.

Nguyen HS, Pham HP, Duong TH, Nguyen TPT et Le HMT. 2016. Personalized Facets for Faceted Search Using Wikipedia Disambiguation and Social Network. Dans : Advanced Computational Methods for Knowledge Engineering. Springer, p. 229-241.

Noll MG et Meinel C. 2007. Web search personalization via social bookmarking and tagging. Dans : The semantic web. Springer, p. 367-380.

O'Donovan J et Smyth B. Trust in recommender systems. Dans : Proceedings of the 10th international conference on Intelligent user interfaces, 2005. ACM, p. 167-174.

Page L, Brin S, Motwani R et Winograd T. 1999. The PageRank citation ranking: Bringing order to the web. Stanford InfoLab.

Paiva S et Ramos-Cabrer M. 2014. The Relevance of Profile-based Disambiguation and Citations in a Fuzzy Algorithm for Semantic Document Search. Procedia Technology, 16 : 22-31.

Pannu M, Anane R et James A. Hybrid profiling in information retrieval. Dans : Computer Supported Cooperative Work in Design (CSCWD), 2013 IEEE 17th International Conference on, 2013. IEEE, p. 84-91.

Pariser E. 2011. The filter bubble: How the new personalized web is changing what we read and how we think. Penguin.

Park LA et Ramamohanarao K. Query expansion using a collection dependent probabilistic latent semantic thesaurus. Dans : Pacific-Asia Conference on Knowledge Discovery and Data Mining, 2007. Springer, p. 224-235.

Pasca M. Towards temporal web search. Dans : Proceedings of the 2008 ACM symposium on Applied computing, 2008. ACM, p. 1117-1121.

PEARL J. 1988. Probabilistic Reasoning in Intelligent Systems. Networks of Plausible Inference.

Pedersen T, Pakhomov SV, Patwardhan S et Chute CG. 2007. Measures of semantic similarity and relatedness in the biomedical domain. Journal of biomedical informatics, 40 : 288-299.

Pepper S et Garshol LM. 2002. The XML Papers: Lessons on Applying Topic Maps. Proceedings of XML 2002 : 1-33.

Pérez-Agüera JR, Arroyo J, Greenberg J, Iglesias JP et Fresno V. Using BM25F for semantic search. Dans : Proceedings of the 3rd international semantic search workshop, 2010. ACM, p. 2.

Petersen T. 1994. Introduction to the Art and Architecture Thesaurus. Oxford University Press.

Phelan D et Kushmerick N. 2002. A descendant-based link analysis algorithm for Web search.

Pollitt AS. MenUSE for medicine: end-user and seraching of MEDLINE via the MeSH thesaurus. Dans : International forum on information and documentation, 1988. International Federation for Information and Documentation, p. 11-17.

Ponte JM et Croft WB. A language modeling approach to information retrieval. Dans : Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval, 1998. ACM, p. 275-281.

Pouliquen B. 2002. Indexation de textes médicaux par extraction de concepts, et ses utilisations. Université Rennes 1.

Preisach C, Marinho LB et Schmidt-Thieme L. Semi-supervised tag recommendation-using untagged resources to mitigate cold-start problems. Dans : Pacific-Asia Conference on Knowledge Discovery and Data Mining, 2010. Springer, p. 348-357.

Prime-Claverie C. 2004. Vers une prise en compte de plusieurs aspects des besoins d'information dans les modèles de la recherche documentaire: Propagation de métadonnées sur le World Wide Web. Ecole Nationale Supérieure des Mines de Saint-Etienne; Université Jean Monnet-Saint-Etienne.

Pujari M et Kanawati R. Tag Recommendation by Link Prediction Based on Supervised Machine Learning. Dans : ICWSM, 2012.

Rada R, Mili H, Bicknell E et Blettner M. 1989. Development and application of a metric on semantic nets. IEEE transactions on systems, man, and cybernetics, 19 : 17-30.

Ralalason B. 2010. Représentation multi-facette des documents pour leur accès sémantique. Université Paul Sabatier-Toulouse III.

Ranganathan SR. 1931. The five laws of library science. Madras Library Association (Madras, India) and Edward Goldston (London, UK).

Rani SG. 2013. A New Ranking Algorithm for Ranking Search Results of Search Engine based on Personalized User Profile. International Journal of Computer Applications, 74.

Rao BB et Vatsavayi VK. 2013. Concept based Ranking of Results using an Ontology and Fuzzy Network for a Personalized Web Search Engine. International Journal of Computer Applications, 81.

Rebaï RZ, Ghorbel L, Zayani CA et Amous I. An adaptive method for user Profile learning. Dans : East European Conference on Advances in Databases and Information Systems, 2013. Springer, p. 126-134.

Resnik P. 1995. Using information content to evaluate semantic similarity in a taxonomy. arXiv preprint cmp-lg/9511007.

Robertson S. 2005. How Okapi came to TREC. Voorhees and Harman (2005) : 287-299.

Robertson S, Zaragoza H et Taylor M. Simple BM25 extension to multiple weighted fields. Dans : Proceedings of the thirteenth ACM international conference on Information and knowledge management, 2004. ACM, p. 42-49.

Robertson SE et Jones KS. 1976. Relevance weighting of search terms. Journal of the American Society for Information Science, 27 : 129-146.

Robertson SE et Walker S. Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. Dans : Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval, 1994. Springer-Verlag New York, Inc., p. 232-241.

Robertson SE, Walker S, Jones S, Hancock-Beaulieu MM et Gatford M. 1995. Okapi at TREC-3. Nist Special Publication Sp, 109 : 109.

Robertson SE, Walker S, Beaulieu M, Gatford M et Payne A. 1996. Okapi at TREC-4. Nist Special Publication Sp : 73-96.

Rocchio JJ. 1971. Relevance feedback in information retrieval.

Rohani VA, Kasirun ZM, Kumar S et Shamshirband S. 2014. An effective recommender algorithm for cold-start problem in academic social networks. Mathematical Problems in Engineering, 2014.

Rose DE et Levinson D. Understanding user goals in web search. Dans : Proceedings of the 13th international conference on World Wide Web, 2004. ACM, p. 13-19.

Rosso P, Ferretti E, Jiménez D et Vidal V. Text categorization and information retrieval using wordnet senses. Dans : Proceedings of the Second International WordNet Conference—GWC, 2004. p. 299-304.

Roxin V et Bernard Y. 2007. Etiquetage collaboratif et nuages de mots: quels apports pour les sites marchands, 6ème Journée Nantaise de Recherche sur le e-Marketing.

Russell S et Norvig P. 2010. Intelligence artificielle: Avec plus de 500 exercices. Pearson Education France.

Ruthven I. Re-examining the potential effectiveness of interactive query expansion. Dans : Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval, 2003. ACM, p. 213-220.

Safi H, Jaoua M et Belguith LH. Intégration du profil utilisateur basé sur les ontologies dans la reformulation des requêtes Arabes. Dans : ACTES DU COLLOQUE, 2015. p. 40.

Safoury L et Salah A. 2013. Exploiting user demographic attributes for solving cold-start problem in recommender system. Lecture Notes on Software Engineering, 1 : 303.

Saha S, Majumder S, Ray S et Mahanti A. Categorizing user interests in recommender systems. Dans : International Conference on Knowledge-Based and Intelligent Information and Engineering Systems, 2010. Springer, p. 282-291.

Said A, De Luca EW et Albayrak S. Inferring contextual user profiles-improving recommender performance. Dans : Proceedings of the 3rd RecSys Workshop on Context-Aware Recommender Systems, 2011.

Salton G. 1971. The SMART retrieval system—experiments in automatic document processing.

Salton G. 1989. Automatic text processing: The transformation, analysis, and retrieval of. Reading: Addison-Wesley.

Salton G et McGill MJ. 1986. Introduction to modern information retrieval.

Salton G et Buckley C. 1997. Improving retrieval performance by relevance feedback. Readings in information retrieval, 24 : 355-363.

Salton G, Fox EA et Wu H. 1983. Extended Boolean information retrieval. Communications of the ACM, 26 : 1022-1036.

Sanderson M. Word sense disambiguation and information retrieval. Dans : Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval, 1994. Springer-Verlag New York, Inc., p. 142-151.

Sanderson M. 2000. Retrieving with good sense. Information retrieval, 2 : 49-69.

Saracevic T. Relevance reconsidered. Dans : Information science: Integration in perspectives In Proceedings of the Second Conference on Conceptions of Library and Information Science, 1996. p. 201-218.

Sarr I, Missaoui R et Lalande R. Dealing with disappearance of an actor set in social networks. Dans : Proceedings of the 2012 International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2012), 2012. IEEE Computer Society, p. 493-500.

Schmetzke A, Greifeneder E et Olson TA. 2007. Utility of a faceted catalog for scholarly research. Library hi tech, 25 : 550-561.

Schwarzkopf E, Heckmann D, Dengler D et Kröner A. Mining the structure of tag spaces for user modeling. Dans : Complete On-Line Proceedings of the Workshop on Data Mining for User Modeling at the 11th International Conference on User Modeling Corfu, Griechenland, 2007. Citeseer, p. 63-75.

Segura A, Vidal-Castro C et Ferreira-Satler M. 2014. Domain Ontology-Based Query Expansion: Relationships Types-Centered Analysis Using Gene Ontology. Dans : Advances in Computational Biology. Springer, p. 183-188.

Seidenberg MS et McClelland JL. 1989. A distributed, developmental model of word recognition and naming. Psychological review, 96 : 523.

Shardanand U et Maes P. Social information filtering: algorithms for automating “word of mouth”. Dans : Proceedings of the SIGCHI conference on Human factors in computing systems, 1995. ACM Press/Addison-Wesley Publishing Co., p. 210-217.

Shaw G, Xu Y et Geva S. 2010. Using association rules to solve the cold-start problem in recommender systems. Advances in Knowledge Discovery and Data Mining : 340-347.

Shen X, Tan B et Zhai C. Context-sensitive information retrieval using implicit feedback. Dans : Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval, 2005. ACM, p. 43-50.

Sieg A, Mobasher B et Burke R. Web search personalization with ontological user profiles. Dans : Proceedings of the sixteenth ACM conference on Conference on information and knowledge management, 2007. ACM, p. 525-534.

Sieg A, Mobasher B, Lytinen S et Burke R. Using concept hierarchies to enhance user queries in web-based information retrieval. Dans : The IASTED international conference on artificial intelligence and applications Innsbruck, Austria, 2004.

Slimani T, BenYaghlane B et Mellouli K. Une extension de mesure de similarité entre les concepts d’une ontologie. Dans : International conference on sciences of electronic, technologies of information and telecommunications, 2007. p. 1-10.

Smith DA et Shadbolt NR. Facetontology: Expressive descriptions of facets in the semantic web. Dans : Joint International Semantic Technology Conference, 2012. Springer, p. 223-238.

Sneha C et Varma G. 2015. USER-BASED COLLABORATIVE-FILTERING RECOMMENDATION.

Song M, Song I-Y, Hu X et Allen RB. 2007. Integration of association rules and ontologies for semantic query expansion. Data & Knowledge Engineering, 63 : 63-75.

Song R, Luo Z, Nie J-Y, Yu Y et Hon H-W. 2009. Identification of ambiguous queries in web search. Information processing & management, 45 : 216-229.

Song X. Enrichment of user profiles across multiple online social networks for volunteerism matching for social enterprise. Dans : Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval, 2014. ACM, p. 1282-1282.

Sorensen H et McElligott M. PSUN: a profiling system for Usenet news. Dans : Proceedings of CIKM, 1995. Citeseer, p. 1-2.

Sparck Jones K. 1972. A statistical interpretation of term specificity and its application in retrieval. Journal of Documentation, 28 : 11-21.

Specia L et Motta E. Integrating folksonomies with the semantic web. Dans : European Semantic Web Conference, 2007. Springer, p. 624-639.

Speretta M et Gauch S. Personalized search based on user search histories. Dans : Web Intelligence, 2005 Proceedings The 2005 IEEE/WIC/ACM International Conference on, 2005. IEEE, p. 622-628.

Stetina J et Nagao M. 1998. General word sense disambiguation method based on a full sentential context. Journal of Natural Language Processing, 5 : 47-74.

Stolze M et Rjaibi W. 2001. Towards scalable scoring for preference-based item recommendation. IEEE Data Eng Bull, 24 : 42-49.

Su X et Khoshgoftaar TM. 2009. A survey of collaborative filtering techniques. Advances in artificial intelligence, 2009 : 4.

Sugiyama K, Hatano K et Yoshikawa M. Adaptive web search based on user profile constructed without any effort from users. Dans : Proceedings of the 13th international conference on World Wide Web, 2004. ACM, p. 675-684.

Sun D, Li C et Luo Z. A content-enhanced approach for cold-start problem in collaborative filtering. Dans : Artificial Intelligence, Management Science and Electronic Commerce (AIMSEC), 2011 2nd International Conference on, 2011. IEEE, p. 4501-4504.

Sy O, Loizou G, Laurent D, Jen T-Y et Jami S. 2016. Extraction de règles d'association pour la prédiction de valeurs manquantes. REVUE AFRICAINE DE LA RECHERCHE EN INFORMATIQUE ET MATHÉMATIQUES APPLIQUÉES, 3.

Tamine-Lechani L, Boughanem M et Zemirli N. Exploiting multi-evidence from multiple user's interests to personalizing information retrieval. Dans : Digital Information Management, 2007 ICDIM'07 2nd International Conference on, 2007. IEEE, p. 7-12.

Tamine-Lechani L, Boughanem M et Zemirli N. 2008. Personalized document ranking: Exploiting evidence from multiple user interests for profiling and retrieval. JDIM, 6 : 354-365.

Tamine L et Calabretto S. 2008. Recherche d'information contextuelle et web. month.

Tamine L, Zemirli N et Bahsoun W. 2007. Approche statistique pour la définition du profil d'un utilisateur de système de recherche d'information. Information-Interaction-Intelligence, 7 : 5-25.

Tchuente D. 2013. Modélisation et dérivation de profils utilisateurs à partir de réseaux sociaux: approche à partir de communautés de réseaux k-égocentriques. Université de Toulouse, Université Toulouse III-Paul Sabatier.

TCHUENTE D, PENINOU A, CANUT M-F, Baptiste-JESSEL N et SEDES F. 2012. Modélisation du processus de développement des profils utilisateurs dans les systèmes d'information.

Thilliez M et Delot T. 2004. Evaluation de requêtes dépendantes de la localisation dans les réseaux mobiles. Premières Journées Francophones: Mobilité et Ubiquité 2004.

Thorat PB, Goudar R et Barve S. 2015. Survey on collaborative filtering, content-based filtering and hybrid recommendation system. International Journal of Computer Applications, 110.

Tika A. 2004. The Apache Software Foundation.

Tomasi F, Ciotti F, Daquino M et Lana M. Using Ontologies as a Faceted Browsing for Heterogeneous Cultural Heritage Collections. Dans : IT@ LIA@ AI* IA, 2015.

Tvarožek M. Personalized navigation in the semantic web. Dans : International Conference on Adaptive Hypermedia and Adaptive Web-Based Systems, 2006. Springer, p. 467-471.

Umbrath ASRWW et Hennig L. 2009. A hybrid PLSA approach for warmer cold start in folksonomy recommendation. Recommender Systems & the Social Web : 10-13.

Uzuner O, Katz B et Yuret D. 1999. Word sense disambiguation for information retrieval. AAAI/IAAI, 985.

Vallet D, Cantador I et Jose JM. Personalizing web search with folksonomy-based user and document profiles. Dans : European Conference on Information Retrieval, 2010. Springer, p. 420-431.

Van Meteren R et Van Someren M. Using content-based filtering for recommendation. Dans : Proceedings of the Machine Learning in the New Information Age: MLnet/ECML2000 Workshop, 2000. p. 47-56.

Vandaele V, Francq P et Delchambre A. Analyse d'hyperliens en vue d'une meilleure description des profils. Dans : Proceedings of JADT, 2004.

Vander T. 2007. Folksonomy coinage and definition.

Vogt CC, Cottrell GW, Belew RK et Bartell BT. Using Relevance to Train a Linear Mixture of Experts. Dans : TREC, 1996. p. 503-515.

Voorhees EM. Using WordNet to disambiguate word senses for text retrieval. Dans : Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval, 1993. ACM, p. 171-180.

Wanaskar U, Vij S et Mukhopadhyay D. 2013. A hybrid web recommendation system based on the improved association rule mining algorithm. arXiv preprint arXiv:13117204.

Wang H, Liang Y, Fu L, Xue G-R et Yu Y. Efficient query expansion for advertisement search. Dans : Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval, 2009. ACM, p. 51-58.

Wang W, Zhang D et Zhou J. 2011. COBA: A Credible and Co-clustering Filterbot for Cold-Start Recommendations. Dans : Practical Applications of Intelligent Systems. Springer, p. 467-476.

Weber I et Castillo C. The demographics of web search. Dans : Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval, 2010. ACM, p. 523-530.

Witten IH, Moffat A et Bell TC. 1999. Managing gigabytes: compressing and indexing documents and images. Morgan Kaufmann.

words C. http://www.k12reader.com/term/compound-words/

Wu Z et Palmer M. Verbs semantics and lexical selection. Dans : Proceedings of the 32nd annual meeting on Association for Computational Linguistics, 1994. Association for Computational Linguistics, p. 133-138.

Xie HI. 2008. Users' evaluation of digital libraries (DLs): Their uses, their criteria, and their assessment. Information processing & management, 44 : 1346-1373.

Xu J. 1997. Solving the word mismatch problem through automatic text analysis. University of Massachusetts Amherst.

Xu S, Bao S, Cao Y et Yu Y. Using social annotations to improve language model for information retrieval. Dans : Proceedings of the sixteenth ACM conference on Conference on information and knowledge management, 2007. ACM, p. 1003-1006.

Xu S, Bao S, Fei B, Su Z et Yu Y. Exploring folksonomy for personalized search. Dans : Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval, 2008. ACM, p. 155-162.

Xue G-R, Han J, Yu Y et Yang Q. 2009. User language model for collaborative personalized search. ACM Transactions on Information Systems (TOIS), 27 : 11.

Yarowsky D. One sense per collocation. Dans : Proceedings of the workshop on Human Language Technology, 1993. Association for Computational Linguistics, p. 266-271.

Yeung A, Man C, Gibbins N et Shadbolt N. 2008. A study of user profile generation from folksonomies.

Yiu ML et Mamoulis N. Efficient processing of top-k dominating queries on multi-dimensional data. Dans : Proceedings of the 33rd international conference on Very large data bases, 2007. VLDB Endowment, p. 483-494.

Ykhlef M et Alqahtani S. 2011. A survey of graphical query languages for XML data. Journal of King Saud University-Computer and Information Sciences, 23 : 59-70.

Yu CT et Salton G. 1976. Precision weighting—an effective automatic indexing method. Journal of the ACM (JACM), 23 : 76-88.

Yujian L et Bo L. 2007. A normalized Levenshtein distance metric. IEEE transactions on pattern analysis and machine intelligence, 29 : 1091-1095.

Zaïer Z. 2010. Modèle multi-agents pour le filtrage collaboratif de l'information. Université du Québec à Montréal.

Zanardi V et Capra L. A scalable tag-based recommender system for new users of the social web. Dans : Database and Expert Systems Applications, 2011. Springer, p. 542-557.

Zayani CA, Péninou A, Canut M-F et Sèdes F. Towards an adaptation of semi-structured document querying. Dans : Held in conjunction with the 6 th International and Interdisciplinary Conference on Modeling and Using Context, 2007. p. 13.

Zemirli N, Tamine-Lechani L et Boughanem M. Présentation et évaluation d'un modèle d'accès personnalisé à l'information basé sur les diagrammes d'influence. Dans : Congrès Informatique des Organisations et Systèmes d'Information et de Décision (INFORSID 2007), 2007. p. 75-86.

Zemirli WN. 2008. Modèle d'accès personnalisé à l'information basé sur les Diagrammes d'Influence intégrant un profil utilisateur évolutif. Université de Toulouse, Université Toulouse III-Paul Sabatier.

Zhai C. 2008. Statistical language models for information retrieval. Synthesis Lectures on Human Language Technologies, 1 : 1-141.

Zhai CX, Cohen WW et Lafferty J. Beyond independent relevance: methods and evaluation metrics for subtopic retrieval. Dans : Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval, 2003. ACM, p. 10-17.

Zhang D, Zou Q et Xiong H. 2013. CRUC: Cold-start recommendations using collaborative filtering in internet of things. arXiv preprint arXiv:13060165.

Zhang M et He C. 2010. Survey on association rules mining algorithms. Dans : Advancing Computing, Communication, Control and Management. Springer, p. 111-118.

Zhang Z-K, Liu C, Zhang Y-C et Zhou T. 2010. Solving the cold-start problem in recommender systems with social tags. EPL (Europhysics Letters), 92 : 28002.

Zhao Q, Hoi SC, Liu T-Y, Bhowmick SS, Lyu MR et Ma W-Y. Time-dependent semantic similarity measure of queries using historical click-through data. Dans : Proceedings of the 15th international conference on World Wide Web, 2006. ACM, p. 543-552.

Zhao S, Du N, Nauerz A, Zhang X, Yuan Q et Fu R. Improved recommendation based on collaborative tagging behaviors. Dans : Proceedings of the 13th international conference on Intelligent user interfaces, 2008. ACM, p. 413-416.

Zheng B, Zhang W et Feng XFB. 2013. A survey of faceted search. Journal of Web engineering, 12 : 041-064.

Zheng N et Li Q. 2011. A recommender system based on tag and time information for social tagging systems. Expert Systems with Applications, 38 : 4575-4587.

Zhou D, Lawless S et Wade V. Web search personalization using social data. Dans : International Conference on Theory and Practice of Digital Libraries, 2012a. Springer, p. 298-310.

Zhou D, Lawless S et Wade V. 2012b. Improving search via personalized query expansion using social media. Information retrieval, 15 : 218-242.

Zhou D, Bian J, Zheng S, Zha H et Giles CL. Exploring social annotations for information retrieval. Dans : Proceedings of the 17th international conference on World Wide Web, 2008. ACM, p. 715-724.

Zhou D, Lawless S, Liu J, Zhang S et Xu Y. Query expansion for personalized cross-language information retrieval. Dans : Semantic and Social Media Adaptation and Personalization (SMAP), 2015 10th International Workshop on, 2015. IEEE, p. 1-5.

Zhou D, Lawless S, Wu X, Zhao W et Liu J. Enhanced personalized search using social data. Dans : Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, 2016. p. 700-710.

Zhou D, Wu X, Zhao W, Lawless S et Liu J. 2017. Query Expansion with Enriched User Profiles for Personalized Search Utilizing Folksonomy Data. IEEE Transactions on knowledge and data engineering.

Zhou K, Yang S-H et Zha H. Functional matrix factorizations for cold-start recommendation. Dans : Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval, 2011. ACM, p. 315-324.

Zhou W, Yu C, Smalheiser N, Torvik V et Hong J. Knowledge-intensive conceptual retrieval and passage extraction of biomedical literature. Dans : Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval, 2007. ACM, p. 655-662.

Zipf GK. 2016. Human behavior and the principle of least effort: An introduction to human ecology. Ravenio Books.

Zubiaga A, García-Plaza AP, Fresno V et Martínez R. Content-based clustering for tag cloud visualization. Dans : Social Network Analysis and Mining, 2009 ASONAM'09 International Conference on Advances in, 2009. IEEE, p. 316-319.