République Algérienne Démocratique et Populaire

Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

Université Yahia Fares de MEDEA

Faculté des Sciences

Département de Mathématiques et d'Informatique

Mémoire de Fin d'Etudes

Pour l'obtention du Diplôme de Master en Informatique

Spécialité : Ingénierie des Systèmes et Technologies du Web

Thème :

Exploitation des Liens d'Amitié de l'Utilisateur pour l'Amélioration des Résultats de Recherche d'Information

Réalisé par :

- MOUAIZIA NesrineImane
- FACI Imene

Encadré par :

- Mr RAFA Tahar

Promotion 2021-2022

Remerciements

On remercie dieu le tout puissant de nous avoir donné la santé et la volanté d'entamer et de terminer ce mémoire.

Tout d'abord, ce travail ne serait pas aussi riche et n'aurait pas pu avoir le jour sans l'aide et l'encadrement de Mr Rafa Tahar, on le remercie pour la qualité de son encadrement exceptionnel, pour sa patience, sa rigueur et sa disponibilité durant notre préparation de ce mémoire.

Nos remerciements s'adressent à Mr Rahim Messaoud pour son aide pratique et ses encouragements, et pour avoir accepté de présider la commission d'évaluation de notre travail.

Nous remercions aussi Mm Sadouki Fatiha d'avoir accepté d'examiner notre travail.

Nos remerciements s'adressentégalement à tous nos enseignants pour leurs générosités et la grande patience dont ils ont fait preuve malgré leurs charges académiques et professionnelles.

Nous remercions Mm Lantri et Mm Abdallah El Hadje enseignantes à l'INSFPde Medea Pour l'aide qu'elles nous ont donné concernant la manipulation du Sql Server

Dédicace

Jr dédie ce mémoire à mes parents, pour l'amour qu'ils m'ont toujours donné, leurs encouragements et toute l'aide qu'ils m'ont apportée durant mes études. Que vous trouverez ici le fruit de tant de dévouements et de sacrifices ainsi que l'expression de ma gratitude et de mon profond amour.

A mes 2 sœurs Yasmine et Sabrine ma source d'espoir et de motivation que dieu vous protège.

A ma chère famille, En témoignage de l'amour et de l'affection qui nous attachent.

Sans oublier les 2 personnes qui était présent pour moi durant c'est 5 ans d'étude de me soutenir et m'encouragé dans tous les épreuves Mokdad Kenza et Laziri Nadji.

Nesrine

Dédicace

A ma chèremère

Qui a attendu avec patience les fruits de sa bonne éducation et de ses dévouements

A celui qui s'est changé la nuit en jour pour m'assurer les bonnes conditions

A mon cher père

A ma chère sœur Hanane et son mari Ali

Qui ont partagé avec moi tous les moments d'émotions lors de la réalisation de ce travail

A mes très chers frères

Qui m'onttoujours soutenu et encouragé durant ces années d'études

A mon cher AkramAmari

Qui m'a donné l'espoir, l'amour et de la vivacité.

À tous les membres de ma famille

Qui m'ont soutenu par meurs prières nuit a jours et leur encouragements

A toutes mes amies

Pourleurs indéfectibles soutiens et leurs patiences infinies.

Imane

RESUME

L'utilisateur d'un système de recherche d'information, de sa nature tant qu'un être humain, est en relation bidirectionnelle d'influence cognitive avec son entourage social. De ce fait, l'utilisateur est souvent intéressé par ce qui intéresse sa communauté proche, et ses besoins en information exprimés par ses requêtes de recherche sont souvent influencés par les sujets qui attirent l'attention de cette communauté. Dans ce mémoire, nous proposons une approche de personnalisation de recherche d'information basée sur la construction d'un profil d'utilisateur. Dans ce profil, nous exploitons les relations sociales de l'utilisateur pour identifier et représenter ses intérêts. Le profil proposé est utilisé à l'enrichissement des requêtes utilisateur pour les aider à bien exprimer ses besoins en information et par conséquent guider ses nouvelles recherches d'information.

Mots-clés : Recherche d'information Personnalisée, profil d'utilisateur, recherche d'information sociale.

ABSTRACT

The user of a social information retrieval system, from its human nature, is in a bidirectional relationship of cognitive influence with his social environment. Thus, the user is often interested by what matters to his close community, and the information need expressed by his search query is often influenced by the topics that attract the attention of this community. In this document, we propose a personalized information retrieval approach based on the construction of a user profile. In this profile, we combine the user search history with the user's social relationships to identify and represent his interests. The proposed profile is used to guide new user information searches by enriching the initial user query.

KEYWORDS:Personalized Information Retrieval, User Profile, Social Information Retrieval.

Tables des Matières

Introduction Générale 2

Chapitre 1 : La recherche d'information classique 6

1. Introduction 6

2. Notion de base 6

3. Processus générale de la recherche d'information 7

3.1. La phase d'indexation 8

3.1.1. L'extraction des mots simples 9

3.1.2. L'élimination des mots vides 9

3.1.3. La normalisation 9

3.1.4. La pondération des termes (TF,IDF) 10

3.2. La Phase de recherche ou d'appariement document-requête 11

4. Les principaux modèles de la RI 11

4.1. Le modèle booléen 12

4.2. Le modèle vectoriel 12

4.3. Le modèle probabiliste 14

5. Evaluation des performances de système RI 14

5.1. Collection de test 15

5.2. Mesures d'évaluation 15

5.2.1. Rappel et Précision 16

5.2.2. F-mesure 17

5.2.3. Précision @X 17

5.2.4. R-Précision 18

5.2.5. La précision moyenne (AP_I) 18

5.2.6. La moyenne des précisions moyennes MAP 18

5.2.7. La moyenne des réciproques des rangs MRR 18

6. Limite de la RI 19

7. Conclusion 19

Chapitre2 : la recherche d'information personnalisée 21

1. Introduction 21

2. La recherche d'information personnalisée 21

3. Modélisation de l'utilisateur 21

3.1. Pertinence Utilisateur 22
3.2. Profil d'utilisateur 22
3.3. Sources des données de profil 23
3.4. Acquisition des données de profil 23
3.4.1. Acquisition explicite 23
3.4.2. Acquisition implicite 24
3.5. Techniques de construction de centres d'intérêts 25
3.6. Représentation des données de profil d'utilisateur 25
3.6.1. Représentation basée sur l'historique de recherche 26
3.6.2. Représentation ensembliste 26
3.6.3. Représentation connexionniste 26
3.6.4. Représentation multidimensionnelle 26
3.6.5. Représentation conceptuelle 27
3.7. Evolution de profil utilisateur 27
4. Les dimensions du contexte de recherche 28
4.1. Dimension sociale 29
.4.2 Dimension localisable 30
4.3. Dimension temporelle 31
4.4. Dimension sémantique 32
5. Intégration de profil utilisateur dans le processus de recherche 33
5.1. Reformulation de requête 33
5.2. Reclassement de résultats 34
6. L'évaluation des systèmes de recherche d'information personnalisés 34
6.1. Les limites de l'évaluation classique 34
6.2. La problématique de l'évaluation personnalisée 35
6.3. Protocole de l'évaluation personnalisée 35
6.4. Les mesures de l'évaluation des systèmes personnalisée 36
7. *Conclusion* 37
Chapitre3 : Exploitation des liens d'amitiés dans un site social pour l'amélioration du résultat de recherche 40
1. *Introduction* 40
2. Profil utilisateur 41
2.1. L'historique de recherche pertinent 42

2.1.1. Les documents pertinents 42
2.1.2. Les termes pertinents 42
2.1.3. Les tags pertinents 42
2.2. La base de liens de cooccurrence *terme-terme* 43
2.3. La base de liens de cooccurrence termes-tags 44
2.4. Liste des amis les plus similaires (pertinence sociale externe) 45
2.5. Mesure de l'influence sociale sur l'utilisateur 46
2.6. Mesure de l'indépendance sociale de l'utilisateur 46
3. Processus de Recherche 47
3.1. Sélection de la base la plus adéquate à la requête d'utilisateur 47
3.2. Reformulation de la requête 48
3.3. Mise à jour de profil 49
4. Evaluation 49
4.1. Difficulté d'évaluation dans le cadre de la recherche personnalisée 49
4.2. Stratégie de l'évaluation 49
5. Conclusion 51
Conclusion Générale 53

Introduction générale

Introduction Générale

La recherche d'information (RI) désigne la branche de l'informatique qui porte sur l'acquisition, l'organisation, le stockage et la recherche de l'information. L'objectif de la RI est de faciliter l'accès à l'information pertinente pour un utilisateur ayant un besoin en information. Le besoin en information de l'utilisateur est souvent exprimé en langage naturel par une requête composée d'un ensemble de mots clés. La requête d'utilisateur est soumise à un système de recherche d'information (SRI) qui permet de retrouver, à partir d'une collection de documents, un sous-ensemble de documents susceptibles d'être pertinents pour la requête d'utilisateur.

Les approches classiques de recherche d'information renvoient, en réponse à une requête utilisateur, un très grand nombre de résultats qui semblent naïvement correspondre à cette requête, même si ces résultats ne sont pas tous importants selon les intentions de l'utilisateur. D'où le besoin d'avoir des connaissances contextuelles sur l'utilisateur pour pouvoir adapter les résultats de recherche à ses intentions, et par conséquent améliorer l'efficacité de la recherche, et c'est ainsi que la recherche d'information personnalisée est apparait .

L'idée principale de la recherche d'information personnalisée, est qu'une même réponse pour une même requête envoyée par deux utilisateurs différents, peut ne pas les satisfaire tous deux. Et ainsi, l'objectif des systèmes de recherche d'information personnalisée est d'inférer le réel besoin en information de l'utilisateur caché derrière sa requête, et de lui retourner par la suite les résultats les mieux correspondants à ce besoin.

Pour pouvoir filtrer les milliers de résultats qui semblent naïvement correspondant à la requête de l'utilisateur, et de classer en premier ceux qui sont les plus pertinents aux intentions de l'utilisateur et non seulement à sa requête, les approches de recherche d'information personnalisée ont besoin de prendre en compte l'utilisateur et les données lui caractérisant au cœur du processus de la recherche d'information. A cet effet, les recherches dans le domaine de la personnalisation se sont principalement concentrées sur la définition d'approches servant à la modélisation de l'utilisateur. Cette modélisation est traduite par une structure appelée dans la littérature de la recherche d'information personnalisée « *profil d'utilisateur* ». Ce profil d'utilisateurregroupe les données que le système de recherche perçoit sur un utilisateur donné. Ce profil sera intégré, par la suite, dans le processus de recherche pour une meilleure adaptation des résultats de recherche aux intentions de l'utilisateur.

L'hypothèse de base des approches de recherche d'information personnalisée est que les nouvelles recherches del'utilisateur sont souvent en relation avec ses recherches précédentes. Et par conséquent, le besoin en information de l'utilisateur peut être inféré en investiguant son historique de recherche. De ce fait, les profils d'utilisateurs ne se limitent pas aux données statiques explicitement exprimées par l'utilisateur telles que l'identité, les domaines d'intérêts, les préférences…etc. mais ils donnent plus d'importance aux données

contextuelles dynamiques qui sont implicitement extraites à partir de l'historique de recherche pertinent de l'utilisateur.

Pour améliorer la représentativité des profils et leur capacité à décrire les intérêts des utilisateurs et d'inférer leurs besoins en information, les approches de la recherche d'information personnalisée exploitent, en plus des données thématiques (documents et termes pertinents), des informations supplémentaires considérées comme des capteurs de pertinences pour l'utilisateur.

Pour certaines approches, l'information sociale (annotations, marquage et partage de document, liens d'amitié, …etc.) est discriminante pour la description de la pertinence utilisateur, de fait qu'elles représentent des avis, des impressions, des jugements, … de l'utilisateur. En plus, les besoins en information de l'utilisateur sont généralement influencés par ce qui intéresse sa communauté proche.

Dans ce travail, nous proposons une approche de personnalisation de la recherche d'information basée sur la construction d'un profil social d'utilisateur. Dans ce profil nous combinons les capteurs de pertinence provenant de :

- Le contexte Thématique, représenté par les termes pertinents extraits à partir des documents pertinents des recherches précédentes de l'utilisateur.
- Le contexte Social interne, représenté par les tags les plus utilisés par l'utilisateur pour annoter les documents pertinents.
- Le contexte social externe, représenté par les liens d'amitié les plus importants reliant l'utilisateur avec les autres utilisateurs de système de recherche.

La contribution de notre approche réside dans :

- l'introduction d'une pseudo-sémantique dans la représentation et l'exploitation des données thématiques et sociales du profil d'utilisateur. Cette pseudo-sémantique s'intéresse à l'identification et la valorisation des différentes relations de cooccurrence reliant les termes pertinents entre eux, et aussi avec les tags les plus fréquents. Notre hypothèse est qu'une apparition fréquente de deux termes (ou un terme et un tag) dans de nombreux documents peut signifier qu'ils coopèrent à la description des mêmes sujets, et que l'un peut avoir une forte probabilité pour désambiguïser l'autre.
- La définition des liens d'amitié entre les utilisateurs de système de recherche relativement à l'importance des intérêts communs entre eux. Cette vision permet de classer en tête de liste des amis d'un utilisateur donné, ceux qui partagent le maximum d'intérêts avec lui, et dont les profils peuvent aider à enrichir la sa requête initial.
- L'introduction d'un score qui mesure la tendance de l'utilisateur à s'intéresser de ce qui intéresse ses amis sur le système de recherche. Notre hypothèse est qu'il n'est pas suffisant de déterminer les amis les plus similaires à l'utilisateur, mais plutôt de déterminer le degré d'influence des amis sur l'utilisateur. Ce score nous permettra de savoir pour un utilisateur donné à quel point il est utile de demander de l'aide auprès des profils de ses amis pour enrichir sa requête.

Notre objectif dans cette approche, est d'exploiter les informations thématiques et sociales fournies par le profil d'utilisateur pour reformuler les requêtes initiales de l'utilisateur dans

le but de les désambigüiser et les aider à mieux exprimer le réel besoin en information de l'utilisateur, et par conséquent pour améliorer la pertinence des résultats de recherche.

Après cette introduction, le reste de document est organisé comme suit :

- Chapitre 1 : Recherche d'information classique
- Chapitre 2 : Recherche d'information personnalisée
- Chapitre 3: L'exploitation des liens d'amitiés dans un site social pour l'amélioration du résultat de recherche

Le document se termine par une conclusion et des perspectives.

Chapitre 01

La recherche d'information classique

Chapitre 1 : La recherche d'information classique

1. Introduction

La recherche d'information est un domaine liée aux sciences de l'information qui permet à l'utilisateur un accès facile à l'information pertinente répondant à son besoin d'information. L'utilisateur formule son besoin en information par une requête en langage naturel transcrite par un ensemble de mots clés. Le système de recherche d'information (SRI) utilise des modèles et des techniques permettant de sélectionner un sous-ensemble de document pertinent à partir d'une collection de documents, afin de satisfaire le besoin d'information de l'utilisateur.

Dans ce chapitre, nous définissons quelques notions de bases de la recherche d'information, en suite nous présentons dans la section 3 le processus générale de la RI qui est répartie en deux phase : phase d'indexation et phase de recherche, ainsi que la section 4 qui aborde les principaux modèles de la RI qui sont classés en trois catégories principales : modèle booléen, modèle vectorielle, modèle probabiliste et pour finir, la section 5 nous mène à l'évaluation des performances de SRI qui décrit la collection de teste et présente les principales mesures de l'évaluation du système de RI. Le chapitre sera conclupar 2 section l'une est une discussion sur les limites de la RI classique, puis l'autre une conclusion.

2. Notion de base

La recherche d'information est traditionnellement considérée comme l'ensemble des techniques permettant de sélectionner à partir d'une collection de documents, ceux qui sont susceptiblesde répondre aux besoins de l'utilisateur. Elle s'intéresse à l'acquisition, l'organisation, le stockage, la recherche et la distribution de l'information.

Plusieurs notions clés sont manipulées dans le domaine de la recherche d'information (RI), nous avons donc trouvé utile de les clarifier :

a) ***Document :***On appelle document toute pièce écrite donnant de renseignement répondant à un besoin d'information d'un utilisateur. un document peut être un texte, un morceau de texte, une image ou une page web …etc.
b) ***Collection de document :***La collection de documents (ou fond documentaire) constitue l'ensemble des informations exploitables et accessibles. Elle est constituée d'un ensemble de documents. Dans le cas générale et pour un souci d'optimalité, la base constitue des représentations simplifiées mais suffisantes pour ces documents.
c) ***Besoin d'information :***est une représentation mentale qui permet à l'utilisateur de s'engagerdans une activité de recherche d'information. Ce besoin est représenté se forme d'une requête. Le besoin d'information est répartir en 3 types :

- ✓ *Besoin Vérificatif :* l'utilisateur cherche à confirmer le texte avec les informations connues qu'il possède déjà.

- ✓ *Besoin Thématique connu :* l'utilisateur cherche à éclaircir de nouvelles informations dans un sujet et domaine connus.
- ✓ *Besoin Thématique inconnue :* l'utilisateur consulte de nouveaux concepts hors sujets ou domaines qui lui sont familiers.

d) ***Requête :***Il s'agit du fait d'interroger un moteur de recherche via la saisie d'un ou plusieurs mots-clésformulés en langage naturel exprimant le besoin d'information de l'utilisateur.

e) ***Pertinence :***la notion de pertinence est le critère le plus important dans la RI et l'un des objets principaux de tous système de recherche d'information, cette notion n'est pas facile à définir car elle fait intervenir plusieurs notions, mais la définition la plus simple est donnée dans [Baeza-Yates,Ribeiro-Neto1999] : « La pertinence est la correspondance entre un document et une requête, ou encore une mesure d'informativité du document à la requête ». il existe plusieurs "'pertinences"' possibles entre un document et un besoin, nous en citons les deux (2) les plus importantes :

- ✓ *pertinence algorithmique (système) :* est traduite par le degré de l'adéquation du contenu du document par rapport à la requête. C'est le seule type de pertinence qui est indépendant du contexte.
- ✓ *Pertinence utilisateur :* pour l'utilisateur la pertinence correspond à la satisfaction d'ensemble des documents restitués par la SRI. Deux utilisateurs peuvent juger différemment un même document renvoyé pour une même requête.

f) ***Le Système de recherche d'information :*** est un ensemble de techniques qui assurent les fonctions nécessaires pour la RI. Son rôle est de retrouver les documents pertinents à une requête d'utilisateur formulé pour répondre à son besoin d'information, à partir d'une base de documents volumineuse et représentée sous une forme particulière (index).

3. Processus générale de la recherche d'information

L'objectif fondamental d'un processus de RI est de sélectionner les documents "les plus correspondant"au besoin de l'information de l'utilisateur décrit par une requête. Pour cela le SRI manipule d'un côté un corpus de documents qu'il transpose à l'aide d'une fonction d'indexation en un corpus indexé et d'un autre coté le besoin d'information exprimés par l'utilisateur sous forme d'une requête qui sera elle-même une requête indexé. Après le système fait la correspondance entre les deux représentations (appariement document-requête) puis retourne une liste de documents estimés pertinents (ce qu'on appelle pertinence Système). La figure 1.1 illustre les deux phases principales de processus de la RI.

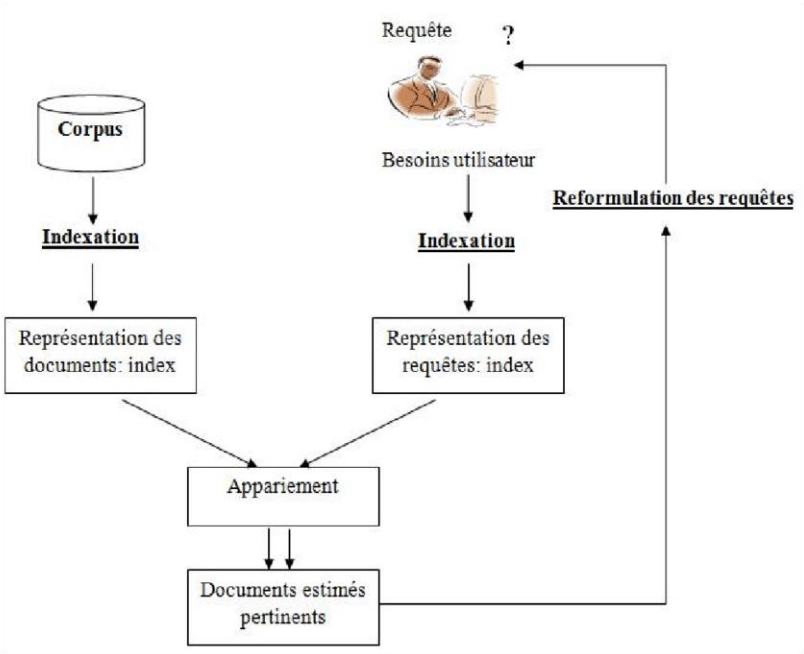


Figure 1.1 Le processus général de la RI

Dans ce qui suit, nous allons citer les deux phases de recherche.

3.1. La phase d'indexation

L'indexation est une étape très importante dans le processus de la RI qui doit s'effectuer avant l'étape de recherche effective de l'information. Cettepremière tâche consiste à analyser les documents et la requêted'utilisateur afin d'extraire les termes représentatifs de leur contenu et construire les index qui faciliteront le processus de recherche, car la qualité de recherche dépend en grande partie de la qualité de l'indexation. Le résultat d'indexation forme ce que l'on appelle le descripteur du document ou de la requête. Le descripteur des documents (mots, groupe de mots) sont rangés dans un catalogue appelée dictionnaire constituant le langage d'indexation.

Différents modes d'indexation existent en RI : l'indexation manuelle, automatique ou semi-automatique

- ✓ *Indexation Manuelle* : dans ce cas, chaque document est analysé par un spécialiste du domaine ou par un documentaliste. Après la lecture des documents, ce spécialiste détermine selon ses connaissances, les mots-clés qui lui semblent les plus adéquats pour représenter le contenu du document. Ce mode d'indexation est fondé sur le jugement humain, il se caractérise par sa profondeur, sa cohérence et sa qualité. [ABBAS 2004]
  L'indexation manuelle à l'avantage d'assurer une meilleure correspondance entre les documents et les termes choisis par les indexeurs pour les représenter, mais l'inconvénient majeur de cette méthode d'indexation est que des termes différents peuvent être sélectionnés par des

indexeurs différents. Il peut même arriver qu'une personne, à des moments différents, indexe différemment le même document. [BAZIZ 2005]

- ✓ *Indexation Automatique* : ce type d'indexation est actuellement le plus répandu car il détecte automatiquement les termes les plus représentatifs du contenu du document, il comprend un ensemble de traitement sur les documents.
  On y distingue : l'analyse de texte du document mot à mot, extraire les mots vides qui ne jouent qu'un rôle syntaxique, éliminer les mots qui apparaissent trop souvent et qui n'ont aucun intérêt, pondérer les termes et finalement la création de l'index.
- ✓ *Semi-automatique* : appelée aussi l'indexation supervisée. C'est un rapport qui combine les deux types d'indexation manuelle et automatique, ainsi l'indexation automatique fait les premiers éléments de l'indexation, puis les spécialistes du domaine corrigent manuellement les informations sélectionnés par l'indexation automatique.

3.1.1. L'extraction des mots simples

L'extraction est l'étape qui permet d'extraire un ensemble de termes ou de mots simples à partir d'un document par une analyse lexicale qui consiste à transformer un document textuel en un ensemble de termes. Pendant cette phase, la ponctuation et la mise en page sont supprimées.

3.1.2. L'élimination des mots vides

Cette opération n'est pas toujours obligatoire. La liste des termes extrait précédemment contient des mots qui ont un rôle syntaxique, non signifiant appelés « mots vides »(appelés aussi : Stop-list ou : anti-lexique)[Fox 1992], tel que :les pronoms personnels, les prépositions, …, etc. L'élimination se fait en utilisant un anti-dictionnaire de mot vide ou en négligeant les termes dépassant un certain nombre d'occurrences dans la collection. L'élimination de mots vides permet de réduire la taille de l'index.

3.1.3. La normalisation

C'est un processus morphologique permettant de regrouper les variantes d'un mot qui ont la même racine en un seul mot appelé : lemme ou racine. Plusieurs techniques sont appliquées pour obtenir les racines des mots, en se basant sur l'élimination de terminaisons. La normalisation se base sur deux méthodes principales : La lemmatisation et la racinisation.

- ✓ *La lemmatisation :* regroupe les différentes formes des mots de la même famille. Ce processus consiste à représenter les mots (ou «lemme ») sous leur forme canonique. Par exemple pour un verbe, ce sera son infinitif. Pour un nom, son masculin singulier. L'idée étant encore une fois de ne conserver que le sens des mots utilisés dans le corpus. On distingue plusieurs types de lemmatisation : la table de

correspondance, l'analyse grammaticale, et utilisation des N-grammes [Adamson et Boreham1974].

- ✓ *La racinisation :* est un processus qui consiste à ne conserver que les racines des mots étudiés. L'idée étant de supprimer les suffixes, préfixes des termes afin de ne conserver que l'origine. C'est un procédé simple que la lemmatisation et plus rapide car on tronque les mots essentiels contrairement à la lemmatisation qui nécessite d'utiliser un dictionnaire.

3.1.4. La pondération des termes (TF,IDF)

La pondération est un élément primordial dans un système de recherche d'information [SparckJones 1974]. Elle permet d'affecter à chaque terme son degré d'importance dans le document, cette importance est calculée sur la base d'aspects statistique. Parmi plusieurs méthodes de pondération dans le domaine, la mesure *TF.IDF* est basée sur la combinaison des deux facteurs : La fréquence du terme (TF), et la fréquence inverse de document (IDF).

- ✓ *La Fréquence du terme (TF-termfrequency) :*

Elle détermine la Fréquence relative d'un mot ou d'une combinaison de mots dans un document. Cette mesure est fondée sur l'idée que plus un terme est fréquent dans un document, plus il est important dans sa description. Elle représente une "pondération locale" d'un terme dans un document. La mesure (TF) peut être présentée par :

- des valeurs binaires : dans ce cas, le $TF_{d,t}$ prends des valeurs binaires.

$$TF_{d,t} = \begin{cases} 1, & si\ le\ terme\ tapparait\ au\ moins\ une\ fois\ au\ document\ d \\ 0, & si\ le\ terme\ t\ n'apparaitpas\ au\ document\ d \end{cases}$$

- Par le nombre d'occurrences (la valeur brut) : Dans ce cas, le $TF_{d,t}$ d'un terme t dans un document d, est le nombre d'occurrences $tf_{d,t}$ de ce terme dans ce document.

$$TF_{d,t} = tf_{d,t}$$

- Par une valeur logarithmique : la valeur de $TF_{d,t}$ est calculée par une fonction logarithmique.

$$TF_{d,t}=1 + \log(tf_{d,t})$$

Où : $tf_{d,t}$ est la valeur brut (le nombre d'occurrence du terme t dans un document)

- Par une valeur normalisée : une fonction de normalisation est appliquée à $TF_{d,t}$.

$$TF_{d,t}= 0.5 + 0.5 \frac{tf_{d,t}}{MaxTF_{t \in d}\,(tf_{d,t})}$$

Où : $MaxTF_{t\in d}$ ($tf_{d,t}$)représente le nombre maximum d'occurrences d'un terme dans un document d

- ✓ *La Fréquence inverse du document (IDF- inverse document Frequency) :*

C'est une pondération globale. Elle calcule la fréquence inverse du document et complète l'analyse de l'évaluation du mot. Il agit en tant que correctif à tous les documents contenant le mot en question, elle est exprimée :

$$IDF_t = Log\ (N/df_t\ )$$

$$IDF_t = Log\ ((N - df_t)/df_t)$$

Où df_t est le nombre de documents contenant le terme t et N est le nombre de documents dans toute la collection. Ils existent d'autres méthodes de pondération comme : La pondération Okapi BM25 [Robertson et al 1996], et la pondération basée sur le model BM27F [Robertson et al 2004][Zaragoza et al 2004].

3.2. La Phase de recherche ou d'appariement document-requête

La phase de recherche implique un processus d'interaction de l'utilisateur avec le SRI. Cette interaction comprend : (1) la formulation d'une requête par l'utilisateur traduisant son besoin en information, (2) la représentation de la requêtesous forme interne selon le mécanisme d'indexation adopté, (3) La correspondance entre la requête et les documents par exploitation de l'index et l'affichage des résultats de recherche. La comparaison entre le document et la requête revient à calculer un score représentatif de la ressemblance entre le document et la requête.se score de similarité entre le document est donné par une fonction nommée *RetrievalStatus Value***.** Elle est notée RSV (d,q), ou d est un document et q est une requête. Le système de recherche retourne à l'utilisateur une liste de documents classés par RSV.

Un modèle de recherche d'information est identifié est caractérisé par la méthode d'indexation des documents et des requêtes et la définition de la fonction de correspondance. Plusieurs modèles de RI ont été développés dans la littérature. Dans la suite, nous présentons les principaux modèles développés dans les travaux dans ce domaine.

4. Les principaux modèles de la RI

Les modèles de la RI sont des programmes qui aident les utilisateurs à trouver les informations qu'ils recherchent dans une collection de documents. Pour une demande d'information donnée, le but de ces modèles est de retourner un sous-ensemble de documents de la collection qui pourraient contenir l'information recherchée (documents pertinents).

Un modèle de RI a pour rôle de fournir une formalisation du processus de RI. il est défini par un quadruplet (D, Q, F, R(q,d))[Baeza-Yates,Ribeiro-Neto1999]: où

– D est l'ensemble de documents

– Q est l'ensemble de requêtes

– F est le schéma du modèle théorique de représentation des documents et des requêtes

– R(q,d) est la fonction de pertinence du document d à la requête q .

Dans la suite, nous présentons un aperçu sur les principaux modèles de la recherche d'information.

4.1. Le modèle booléen

Le modèlebooléen[Salton1971] est le premier qui s'est imposé dans le monde de la recherche d'information. Il s'est imposé grâce à la simplicité et à la rapidité de sa mise en œuvre. Il est basé sur la théorie des ensembles et l'algèbre de Boole. Le modèle booléen considère que les termes de l'index sont présents ou absents d'un document, en conséquence, les poids des termes dans l'index sont binaires c.à.d.w_{ij}= {0,1}. Chaque document est représenté comme une conjonction logique des termes non pondérés qui le composent. Un exemple de représentation d'un document est comme suit : $d_i = t_1 \wedge t_2 \wedge t_3 \wedge \ldots \wedge t_n$.

Une requête q est représentée par une expression booléenne dont les termes sont reliés par des opérateurs logiques (AND,OR, NOT) permettant d'effectuer des opérations d'insertion, d'union et de déférence entre les ensembles de résultats associes à chaque termes. Un exemple de représentation d'une requête est comme suit : $q = (t_1 \wedge t_2) \vee (t_3 \wedge \neg t_4)$.

La fonction de correspondance est basée sur l'hypothèse de présence/absence des termes de la requête dans le document et vérifie si l'index de chaque document *d* implique l'expression logique de la requête *q*. Autrement dit, un document est sectionné si et seulement s'il satisfait l'expression booléenne. Le résultat de cette fonction est : RSV (q, d) = {1, 0}.

Dans le modèle booléen de base, tous les documents qui satisfont une requête sont retrouvés. Ils ne sont pas classés selon leur pertinence. Cela est dû au fait qu'un document satisfait une requête ou il ne le fait pas. Une approche venant compléter le modèle booléen de base : le modèle booléen étendu[Salton,McGill1983].Ce modèle peut être vu comme combinaison du modèle booléen et vectoriel. Il consiste à associer des poids d'indexation à chaque terme d'une requête et d'un document. Ceci permet au SRI de supporter un appariement approché et de mesurer un score de pertinence requête-document au lieu d'un appariement exact (existence ou absence d'un terme dans le modèle booléen de base).

4.2. Le modèle vectoriel

GerardSalton[Salton1971]et ses collègues ont proposé un modèle basé sur le critère desimilitude de Luhn[Peter 1957]qui a une forte motivation théorique [Salton,McGill1983]. Ils ont considéréles représentations de l'index et la requête en tant que vecteurs incorporés dans unespace euclidien de M dimensions, ces dimensions étant les termes du vocabulaired'indexation, où chaque terme est attribué à une dimension

indépendante. Chaque document est représenté par un vecteur: dj = (w_{1j}, w_{2j}, ..., w_{Mj}). De même chaque requêteest représentée par un vecteur : q_i = (w_{1i}, w_{2i}, ..., w_{Mi}).

Avec : *w* correspond au poidsd'un terme dans le document d_j ou dans la requête q_i. La pondération des composantes de la requête est soit la même que celle utilisée pour les documents, soit donnée parl'utilisateur lors de sa formulation.

- ✓ *la fonction (TF.IDF) :* c'est la fonction de pondération la plus répondue. Le poids d'un terme t_B dans un document d_C est défini comme suit :

$$w_{i,j} = tf_{i,j} \times idf_i$$

Où :

- tf_B, C'est la fréquence d'apparition du terme t_B dans le document d_C. (Le pouvoir descriptif du terme)
- idf_Best la fréquence inverse de documents. (le degré de généralité du terme)

Il existent d'autres méthodes de pondération [Salton et Allan 1994][Salton,McGILL 1983] [Baeza-Yates, Ribeiro-Neto 1999] plus efficaces, qui combinent les deux mesures TF et IDF, et intègrent un facteur de normalisation.

$$T F * IDF = \log (1 + T F) * \log (N/ df_t)$$

La correspondance entre la requête et le document esttraduite par la similarité de leurs vecteurs associés. Le mécanisme de recherche consistedonc à retrouver les vecteurs documents qui s'approchent le plus du vecteur requête.Les principales mesures de similarité utilisées sont :

- *le produit scalaire :*

$$RSV(q_i, d_j) = \sum_{k=1}^{M} w_{ki} \,.\, w_{kj}$$

- *la mesure de Jaccard :*

$$RSV(q_i, d_j) = \frac{\sum_{k=1}^{M} w_{ki}.w_{kj}}{\sum_{k=1}^{M} w_{ki}^2 + \sum_{k=1}^{M} w_{kj}^2 - \sum_{k=1}^{M} w_{ki}.w_{kj}}$$

- *la mesure cosinus :*

$$RSV(q_i, d_j) = \frac{\sum_{k=1}^{M} w_{ki}.w_{kj}}{\sqrt{\sum_{k=1}^{M} w_{ki}^2} \cdot \sqrt{\sum_{k=1}^{M} w_{kj}^2}}$$

Les principaux avantages du modèle vectoriel sont liés tout d'abord, à la pondérationon binaire des termes, ce qui offre une meilleure qualité des résultats. Le modèle permet une correspondance partielle ou approximative entre les documents et les requêtes (best match). Les documents sont triés selon leur degré de similarité vis-à-vis de larequête. La longueur des documents est traitée naturellement dans l'appariement, carelle est considérée

dans le calcul des poids des termes. Cependant, le modèle vectoriela l'inconvénient de considérer l'indépendance de tous les termes de l'index. Par contre,en pratique, la prise en compte globale de la dépendance des termes peut induire àune baisse de qualité des réponses du système car les dépendances sont généralement locales. C'est pour ces raisons que le modèle vectoriel préserve sa popularité enrecherche d'information, et reste encore utilisé comme modèle de référence lors d'uneévaluation de comparaison de modèles

4.3. Le modèle probabiliste

Le modèle probabiliste a été développé dans les années 70, et a connu des développements récents, car les approches basées sur ce modèle ont obtenu de très bons résultats dans TREC. Il se base sur un calcul de similarité probabiliste[Robertson,SparckJones 1976][Salton,McGill1983][Maron, Kuhns1960].La fonction de correspondance entre un document d_C et une requête Q, se base sur le calcul de la probabilité de pertinence entre le document et la requête. Pour calculer cette probabilité, il faut d'abord calculer deux probabilités :

- P (R,d_j): La probabilité que le document d_C soit pertinent pour la requête Q.

- P ($\bar{R}|d_j$) : La probabilité que le document d_C ne soit pas pertinent pour la requête Q.

La similarité entre le document d_C et la requête Q, est calculée par la formule suivante:

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

Où : R est l'évènement de pertinence, $\bar{R}$ est l'évènement de non pertinence.

Ce modèle a donné lieu à de nombreuses extensions. Il est à l'origine du système OKAPI. Le modèle Okapi BM25 a été développé par (Robertson et al.1994) dans lequel le calcul du poids d'un terme dans un document intègre desaspects relatifs à la fréquence locale des termes, leur rareté et la longueur des documents.

5. Evaluation des performances de système RI

La validationd'un SRI se base sur l'évaluation expérimentale de ses performances. Cette évaluation a toujours été un centre d'intérêt important. Pour cela des mesures de qualité des systèmes de recherche ainsi que des ensembles de données ont été développés afin de tester ces systèmes sur une base commune. Laqualité d'un système est mesurée en comparent les réponses d'un système avec les réponses que l'utilisateur espère. Cette évaluation est fondée sur la notion de pertinence des résultats et elle est souvent difficile à définir et qualifier car elle est relative aux utilisateurs.

On ditqu'un document est pertinent s'il répond au besoin d'information de l'utilisateur, on différencie 2 types de pertinences :

- *La pertinence utilisateur :* l'utilisateur juge les documents pertinents par rapport à son besoin d'information.[Bouhini2014]

- *La pertinence système :* le SRI juge la pertinence des documents pour une requête sur la base de sa fonction de correspondance RSV(d,q), entre le document et la requête.[Bouhini2014]

La pertinence utilisateur doit être la plus proche possible de la pertinence système.

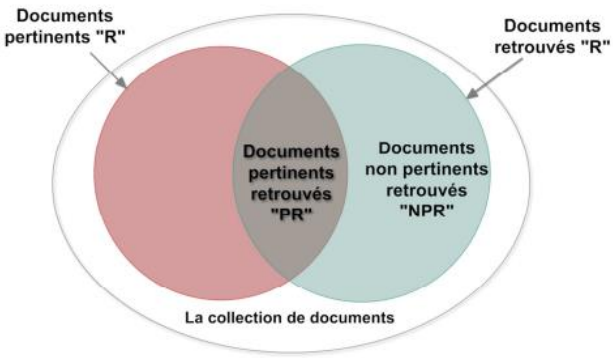


Figure 1.2 Evaluation en recherche d'information

5.1. Collection de test

Pour évaluer le degré de pertinence des réponses d'un système de recherche d'information à une requête,il est nécessaire de connaitre l'ensemble des documents pertinents pour une requête donnée .C'est à cette fin que les collections de test ont été mis au point pour évaluer la performance des SRI. Une collection de test est constitué de :

- ✓ *Collection de document (corpus) :* c'est un ensemble de document indexé sur lesquels le système de recherche d'information sera évalué.[Saoud2017]
- ✓ *Collection de requêtes :*

  Contient plusieurs requête, c'est une représentation d'un besoin d'information formuler par une liste d'un ou plusieurs mot clés.[Saoud 2017]
- ✓ *Jugements de pertinence:*

  Ils sont manuellement établis par l'utilisateur lui-même ou par un observateur expert extérieur, ils permettent d'indiquer la liste de documents pertinents pour chaque requête. Parfois les jugements de pertinence indiquent le degré de pertinence entre les documents et les requêtes.

Pour évaluer la performance d'un SRI, il faut comparer pour chaque requête, les résultats retournés par le SRI aux résultats jugés pertinents dans la collection. Parmi les collections de test les plus utilisées dans le domaine de recherche d'information, on trouve les collections fournis par la compagne TREC, les collections fournis par la compagne INEX, Les collections fournis par la compagne NTCIR, et les collections fournies par la compagne CLEF.[Saoud 2017]

5.2. Mesures d'évaluation

En RI, la mise au point des modèles de recherche passe par une phase expérimentale qui apour but de comparer entre les modèles ou la mise au point de leurs paramètres.

L'enjeu du processus de recherche d'information est de minimiser la distance entre la pertinence du système et celle l'utilisateur. Pour évaluer la performance d'un SRI, deux principales mesures statistiques, entre autres, sont utilisées : la précision et le rappel.

5.2.1. Rappel et Précision

Le rappel et la précision sont deux mesures pour évaluer les performances des systèmes de recherche d'information, la précision est la proportion des items pertinents parmi l'ensemble des items proposés et le rappel est la proportion des items pertinents proposés parmi l'ensemble des items pertinents. Ces deux notions correspondant ainsi à une conception et à une mesure de la pertinence.

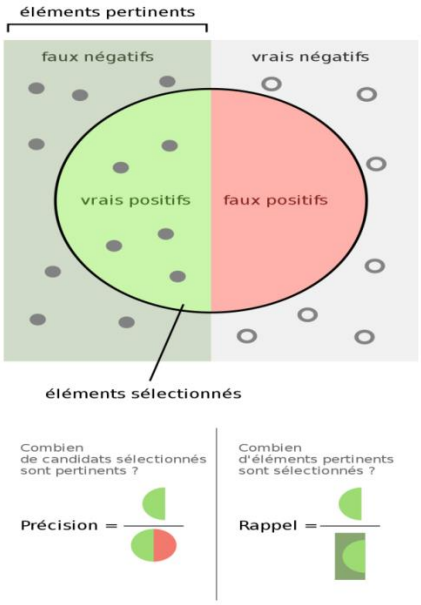


Figure 1.3 Précision-Rappel

✓ ***Le Rappel :***

Cette notion est vue comme une mesure de couverture du système. Le rappel est défini par le nombre de documents pertinents retrouvées au regard du nombre de documents pertinents que possède la base de données. Il est calculé comme suit

$$\text{Rappel (R)} = \frac{\text{vrai positifs}}{\text{vrai positif + faux négatifs}} = \frac{\text{Nombre de Document pertinents selectionnés}}{\text{Nombre de tous les Documents Pertinents}}$$

En statistique, le rappel est appelé sensibilité.

✓ ***La Précision :***

Cette notion calcule la capacité du SRI à retrouver uniquement les documents pertinents. La précision est le nombre de documents pertinents retrouvés rapporté au nombre documents total proposé pour une requête donnée. On calcule la précision avec la formule suivante :

$$\text{Précision(P)}=\frac{\text{vrai positifs}}{\text{vrai positif +faux positifs}}=\frac{\text{Nombre de Document pertinents selectionnés}}{\text{Nombre de Documents selectionnés}}$$

En statistique, la précision est appelée valeur prédictive positive.

Le Rappel n'est pas une mesure suffisante, car il ne montre pas qu'il ya des documents non pertinents parmi les résultats. Ces deux mesures sont dépendantes l'une de l'autre. Si on augmente le rappel en retrouvant plus de documents pertinents, et si on diminue la précision en retrouvant aussi plus de documents non pertinents. Inversement, une plus grande précision risque de rejeter des documents pertinents diminuant ainsi le rappel.[Boubekeur2008]

La courbe de Précision-Rappel à l'allure générale présentée dans la Figure suivante :

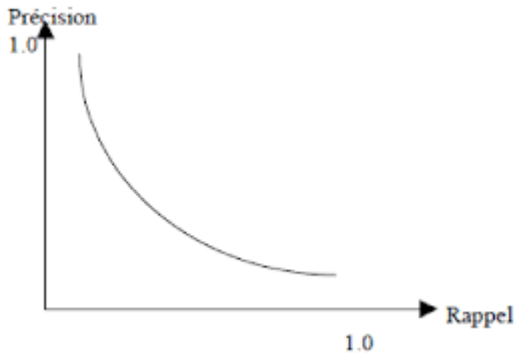


Figure 1.4 La courbe Précision-Rappel

Deux autres mesures complémentaires à la précision et au rappel sont définies :

- ***Bruit*** : est une mesure complémentaire à la *Précision*. Le *bruit* est définit comme *B=1-P*. Il représente le nombre de documents non pertinents retournés par le SRI.
- ***Silence*** : est une mesure complémentaire du *Rappel*. Le *Silence* est définit comme *C=1-R*. Il représente le nombre de documents pertinents non retournés par le SRI.
-

5.2.2. F-mesure

Une mesure qui combine la précision et le rappel est leur moyenne harmonique, nommée F-mesure ou F-score :

$$F = \frac{2 * Precision * Rappel}{Precision + Rappel}$$

5.2.3. Précision @X

C'est la précision à différents niveaux de la liste des documents retournés par le système. Cette précision mesure la proportion des documents pertinents retrouvés parmi les X premiers documents retournés par le système.

Par exemple, P@10 est la proportion des documents pertinents restitués parmi les 10 premiers documents retrouvés.

5.2.4. R-Précision

Cette métrique nécessite de connaitre tous les documents pertinents pour une requête. Le nombre de documents pertinents, R est utilisé comme limite pour le calcul, et cela varie d'une requête à l'autre.

5.2.5. La précision moyenne (AP_I)

La précision moyenne pour AP_I (Averageprecision) est calculée pour chaque requête q_i. Il s'agit de la moyenne des Précisions à chaque rang de document pertinent pour cette requête. Elle est donnée par la formule suivante :

$$AP_I = \frac{\sum_{k=1}^{N} rel(k) \times p(k)}{\sum_{k=1}^{N} rel(k)}$$

Avec :

N : le nombre de document retourné par le SRI.
P(k) : précision du k^{ième} document.
Rel(k) : vaut 1 si le k^{ième} document est pertinent, 0 sinon.

5.2.6. La moyenne des précisions moyennes MAP

Cette mesure permet de mesurer la performance globale du système. Il s'agit de la moyenne des précisions moyennes (MAP) pour l'ensemble des requêtes. Elle est calculée à l'aide des précisions moyennes AP des requêtes, comme suit :

$$\boldsymbol{MAP = \frac{\sum_{q \in Q} AP_q}{|Q|}}$$

Où |Q| est le nombre total de requêtes.

5.2.7. La moyenne des réciproques des rangs MRR

Cette mesure (*MeanReciprocal Rank MRR*) permet d'évaluer le nombre de documents qu'il faut considérer avant de retrouver le premier document pertinent. Elle est égale à la moyenne (calculée sur l'ensemble des requêtes) du rang du premier document pertinent. Cette mesure est donnée par la formule suivante :

$$MRR = \frac{1}{|q|} \sum_{i=1}^{|q|} \frac{1}{rank\ (i)}$$

Où *Rank(i)* est le rang du premier document pertinent pour la i^{eme} requête.

6. Limite de la RI

La recherche d'information classique se base principalement sur l'appariement des descripteurs des documents et des requêtes pour la sélection des informations pertinenteset qui répond au besoin en information spécifique de l'utilisateur.

La principale limitation de la RI classique est qu'elle repose sur une approche générale qui évalue toujours les requêtes des utilisateurs et fournit des résultats, quels que soient les critères spécifiques ou le contexte de recherche de l'utilisateur qui a émis la requête. [Daoud 2009].

Des recherches ont montré que la source de ces limitations se trouve en partie dans le modèle de recherche classique qui considère que les besoins d'information d'un utilisateur doivent être entièrement représentés par sa requête, puis fournit des résultats en ne considérant que les critères de sélection de contenu et la disponibilité des sources d'information [Buidguaguen 2011]. D'autres facteurs viennent accentuer ce problème, notamment : des requêtes utilisateurs courtes, des volumes D'informations importants, des expressions plus ou moins vagues du besoin en Information, ont donné naissance à la recherche d'information personnalisée qui pour objectif de tenir compte, lors de la recherche d'information, de l'utilisateur et de ses intérêts pour améliorer la pertinence des résultats de recherche.

7. Conclusion

Au cours de ce chapitre nous avons présenté les concepts de base de la RI classique ainsi une aperçu sur son historique, les mots clés manipulés, le processus d'accès à l'information, les différents composants et étapes d'un système de recherche ainsi que les principaux modèles de recherche. En s'achevant à la phase d'évaluation des performances des systèmes de recherche d'information.

Compte tenu des limitations de la RI classique les travaux se sont dirigées vers la RI personnalisées, dont le but de répondre mieux au besoin en information de l'utilisateur en améliorant la pertinence des résultats de recherches.

Le chapitre suivant nous mène à la recherche d'information personnalisé.

Chapitre 02

La recherche d'information personnalisée

Chapitre2 : la recherche d'information personnalisée

1. Introduction

Vue la croissance continue du nombre et des types des documents disponibles dans le web, il devient de plus en plus difficile pour l'utilisateur de trouver un contenu pertinent pour sa requête. Les moteurs de recherche disponibles renvoient habituellement des milliers de résultats par question [Zemirli 2008]. Ces résultats ne sont pas tous pertinents pour l'utilisateur et ne satisfont pas le besoin en information de l'utilisateur. Ce qui a poussé Plusieurs travaux à se concentré sur l'amélioration d'accès à l'information.

La personnalisation est une dimension permettant l'implémentation de systèmecentrés utilisateur afin d'approprié son fonctionnement au contexte exact de l'utilisateur.

La recherche d'information personnalisée essaye de se focaliser sur la modélisation de l'utilisateur selon un profil qui regroupe l'ensemble d'informations caractérisant l'utilisateur et ses intérêts, puis l'intégrer dans le processus d'accès aux contenus, dans l'intention de mieux répondre à ses besoins spécifiques.

2. La recherche d'information personnalisée

La personnalisation est une dimension qui permet la mise en œuvre de systèmes centrés utilisateurs, en vue d'adapter son fonctionnement a son contexte précis [zemirli 2008]. Elle ne se limite pas seulement à modéliser les caractéristiques des utilisateurs des profils mais de déduire à partir de ces profils, l'intention de l'utilisateur lorsqu'il effectue sa recherche, en d'autre terme son contexte de recherche. La recherche personnaliséeexploite le profil d'utilisateur pour lui fournir une information pertinente correspondant à ces préférences et à ses besoins [Abdou et al 2006].

Pour kostadinov[Kostadinov 2003], la personnalisation de l'information se définit par un ensemble de préférences individuelles, par des ordonnancements de critères ou par des règles sémantiques spécifiques à chaque utilisateur ou communauté d'utilisateurs. Ces modes de spécification servent à décrire le centre d'intérêt de l'utilisateur, le niveau de qualité des données qu'il désire ou des modalités de présentation de ces données.

3. Modélisation de l'utilisateur

La modélisation de l'utilisateur est une discipline de recherche dont la principale préoccupation est d'améliorer la qualité des interactions homme-machine par déduction et prédiction des buts, préférences et contexte des utilisateurs à partir des faits observés [Haddadlakri 2015]

Le profil de l'utilisateur, qui représente la concrétisation de sa modélisation, couvre des aspects larges tels que son environnement cognitif, social et professionnel représentant des intérêts au cours d'une session de recherche ; il contient généralement des informations (mots-clés, concepts, annotations, dates, location, URLs, …etc). Ce profil utilisateur construit est évalué et utilisé dans des applications spécifiques telles que les systèmes de

recommandation ou de recherche d'information personnalisée. Les approches et techniques de la modélisation de l'utilisateur sont basées sur des modèles simples ou complexes en fonction du but final ou le domaine d'applications du système.

On peut résumer la modélisation de l'utilisateur par la question suivante: quelles sont les informations significatives sur l'utilisateur pouvant représenter des capteurs de pertinence pour lui, que doit contenir son profil pour guider ses nouvelles recherches?

3.1. Pertinence Utilisateur

Le but fondamental de tout SRI est de retourner une information pertinente et ignorer celle non pertinente. La pertinence utilisateur est une notion subjective dépend du point de vue de l'utilisateurlui permettant d'exprimer sa satisfaction par rapport aux documents que les systèmes lui restitue.

La pertinence utilisateur est définit à travers des critères varient entre la subjectivité et l'objectivité qui sont liées au contexte de la recherche. En effet, deux utilisateur ayant entrée la même requête au système, ne jugent pas les réponses de la même manière et cela est due aux différents besoins en information des deux utilisateurs.

La modélisation de l'utilisateur, à travers son profil, permet de comprendre ses mécanismes cognitifs, notamment ceux permettant de percevoir le concept subjectif de la pertinence. Ainsi, un modèle d'utilisateur doit mettre en évidence toute information pouvant aider à décrire ou à inférer la pertinence utilisateur.

3.2. Profil d'utilisateur

Afin de personnaliser l'information, le système de recherche d'information personnalisé doit contenir des éléments clés ayant une incidence concrète sur la recherche en cours. Cet élément représente les données inclus dans le profil d'utilisateur et les métadonnées des documents.

Un profil utilisateur est défini comme « une source de connaissance qui contient des acquisitions qui peuvent être utiles pour le comportement du système. »[whlster 1986]

Le profil d'utilisateur contient : (les données personnelles, l'historique/feedbacks, les annotations associées aux documents, les préférences et les intérêts) ; les données personnelles sont stables et ne demandent pas de mise à jour automatique contrairementaux préférences et les intérêts qui changent au fil du temps. Il permet au système de s'adapter à l'utilisateur.

Pour modéliser le profil, les questions à poser sont : comment les préférences d'utilisateur peuvent être obtenues et comment seront-elles structurées.

En effet, le processus de modélisation du profil utilisateur requiert deux étapes [Amato 1999]. Il faut déterminer en premier le « Quoi » puis le « Comment » :

- ✓ Le «Quoi » : c'est la détermination de ce que doit représenter le contenu du profil. Quelles sont les informations pertinentes représentant au mieux les centres d'intérêts et les besoins de l'utilisateur ?

- ✓ Le « Comment » : c'est la détermination de la structure du profil et la/les technique(s) à utiliser pour la construction de ce profil.

3.3. Sources des données de profil

L'importance des informations utilisées pour la modélisation des utilisateurs influence directement à la présentation et les qualités des profils finaux. Nousdistinguons trois types et sources d'information :

- ✓ *Exploitation des contenus générés par l'utilisateur :*

  Afin d'extraire les centres d'intérêts des utilisateurs, il faut récolter des informations sur les activités des utilisateurs par exemple les documents visités, les pages web et site favoris, exploiter son environnement (temps location) et des informations locales et contextuelle.

- ✓ *Exploitation des relations sociales des utilisateurs*:

  Les informations provenant des réseaux sociaux tels que les tweets, les commentaires, les publications les tags, et le cercle sociale des utilisateurs (liens d'amitiés). Aide à déduire les centres d'intérêts des utilisateurs et enrichir les profils d'utilisateurs.

- ✓ *Exploitation des données externes* :

  Les données externes telles que le contenu URLs, les documents annotés par les utilisateurs, ou l'agrégation de tous leurs comptes sur les différentes plateformes. Elles fournissent une meilleure modélisation des centres d'intérêts des utilisateurs.

L'analyse des données externes exige un effort et n'est pas toujours disponible. De plus, elle peut introduire un bruit (sélectionné un contenu non pertinent).

3.4. Acquisition des données de profil

L'acquisition et la collecte des données des utilisateurs est une étape très importante dans le développement d'un système personnalisé de recherche d'informations. L'acquisition de données utilisateur peut être effectuée implicitement ou explicitement. Dans la récupération explicite, les données sont directement récupérées auprès de l'utilisateur lui-même. Dans l'acquisition implicite, les données de l'utilisateur sont acquises à partir de son comportement, de ses interactions avec le système et de son historique de recherche, et cela sans l'intervention directe de l'utilisateur

3.4.1. Acquisition explicite

L'acquisition explicitepermet aux utilisateurs d'introduire manuellement leurs centres d'intérêts et leurs préférences. Les utilisateurs peuvent être interrogés directement à travers des interfaces des sites web. Dans d'autres cas, l'acquisition explicite utilise des techniques de feedback explicite largement utilisées dans les systèmes de filtrage et de reformulation de requêtes par réinjection de pertinence [Daoud 2009, Saoud 2017]. Ces

techniques sont utilisées dans le cadre de la recherche d'information personnalisée dans le but de maintenir et faire évoluer le profil de l'utilisateur.

Ces données explicites peuvent comprendre : des données d'identité (identifiant d'utilisateur, nom, prénom, etc.), des données démographiques (âge, sexe, adresse, etc.), des caractéristiques physiques (taille, couleur des yeux, etc.), le parcours scolaire, l'emploi occupé, etc. Nous pouvons également ajouter à cette catégorie des données explicites ou des jugements (commentaires explicites) exprimés par les utilisateurs. Par exemple, fournir des évaluations (les évaluations des produits achetés sur Internet par les internautes), effectuer des actions recommandées (par exemple, des articles que les utilisateurs recommandent à d'autres utilisateurs) dans une plage de valeurs prédéfinies. Les données explicites sont des éléments d'un profil utilisateur directement exploitables par les mécanismes de personnalisation [Tchuente 2013].

Ce mode explicite de collecte de données se caractérise par sa mise en œuvre simple et sa structure maîtrisée des profils utilisateurs, mais il présente des limites. D'une part, la récupération explicite n'améliore pas efficacement la représentation du modèle utilisateur lorsque les propriétés de l'interface de récupération explicite ne sont pas bien garanties. D'autre part, l'effort supplémentaire imposé aux utilisateurs pour spécifier explicitement leurs besoins est indésirable pour les utilisateurs du réseau et coûteux à exécuter.

Ces limitations ont orienté les travaux vers destechniques d'acquisition implicite des données du contexte utilisateur.

3.4.2. Acquisition implicite

Consisteà collecter, à l'aide d'algorithmes d'acquisition implicite, les données de l'utilisateur en observant ses interactions avec le système durant les activités de recherche [Buidguaguen 2011].L'acquisition implicite de données utilisateur a pour objectif principal de décharger l'utilisateur de la contrainte de saisie d'informations sur ses centres d'intérêts. Cette acquisition est effectuée au cours de la navigation et des interactions de l'utilisateur avec le système de recherche. Les comportements de navigation sont observés, enregistrés et ensuite traités.

Afin de construire le centre d'intérêt d'un utilisateur à partir de son comportement, l'interprétation de ces comportements nécessite l'utilisation de métriques qui reflètent l'intérêt de l'utilisateur à manipuler des données. La définition et la sélection de ces métriques constituent le principal problème d'identification des centres d'intérêt des utilisateurs à partir de données implicites. La plupart des travaux sur la rétroaction implicite se sont concentrés sur la définition de métriques permettant de bonnes prédictions sur la pertinence du comportement observé par l'utilisateur [Tchuente 2013].

Parmi les interactions des utilisateurs observables dans l'historique de navigation pouvant représenter des indicateurs de pertinence, on cite :

- Les interactions relatives à la lecture (examiner, défiler, regarder, écouter,…etc.) d'un document, d'une page web, d'une séquence vidéo ou une autre source

multimédia. Les indicateurs sont (durée de regarder, durée de défilement, nombre de clics de souris ou des touches de défilement clavier, sélections de textes, ….etc.)

- Les interactions post-lecture : télécharger, imprimer, partager, annoter, commenter, recommander, …etc. une source web.

D'autres interactions peuvent être observées à l'aide de dispositifs technologiques externes, tels que le suivi des mouvements oculaires sur l'écran, l'analyse de la fréquence cardiaque lors des interactions de l'utilisateur, l'interprétation des expressions faciales de l'utilisateur, etc.

Le principal avantage de cette méthode implicite est qu'elle ne nécessite aucune implication directe de l'utilisateur, mais l'inconvénient est qu'elle n'est pas aussi précise que l'évaluation explicite de l'utilisateur, mais elle est plus dynamique dans le temps.

3.5. Techniques de construction de centres d'intérêts

Les modèles des centres d'intérêts des utilisateurs sont construits en utilisant une variété de techniques d'extraction d'informations à partir des sources d'information utilisateur [Krir 2012]:

- *Extraction d'ensembles de termes* : La technique d'extraction d'ensembles de termes est basée sur des techniques d'analyse statistique de mots-clés. En effet, le contenu des documents consultés par les utilisateurs est analysé pour en extraire les mots clés importants. Les termes extraits sont pondérés par des poids reflétant leur importance, dans le but de former des vecteurs de termes qui représentent le centre d'intérêt de l'utilisateur.
- *Extraction de réseau de termes* : Semblable à la technique d'extraction d'ensembles de termes, les termes sont extraits des documents que l'utilisateur considère comme pertinents. Cependant, la différence réside dans la terminologie exprimée en tant que réseau de nœuds. Cette technique implique principalement la représentation sémantique du modèle utilisateur. Afin de construire un modèle utilisateur, il est nécessaire de tirer parti des relations préexistantes entre les termes et les concepts. Ces relations peuvent être trouvées dans des sources sémantiques telles que WordNet.
- *Extraction de concepts* : Cette technique de construction implique principalement un modèle utilisateur représenté par une hiérarchie de concepts pondérée. La raison d'être de cette technique est d'utiliser une taxonomie de concepts de référence comme profil de base. La méthode de construction présente généralement les deux étapes suivantes : (1) identifier le concept et la hiérarchie de l'ontologie à utiliser ; (2) extraire le centre d'intérêt de l'utilisateur par analogie aux concepts de l'ontologie.

3.6. Représentation des données de profil d'utilisateur

Le profil utilisateur peut être représenté selon plusieurs modèles de représentations. Nous détaillons dans ce qui suit les principaux modèles de représentation de profile utilisateur.

3.6.1. Représentation basée sur l'historique de recherche

Le profil utilisateur peut être représenté selon plusieurs modèles de représentations. Nous détaillons dans ce qui suit les principaux modèles de représentation de profile utilisateur.

3.6.2. Représentation ensembliste

Elle se base sur un ensemble de mots clés (ou vecteurs de termes) pondérés représentés souvent selon le modèle vectoriel de Salton.Le poids d'un terme est souvent calculé selon le schéma TF*IDF. Les paquets de termes représentent généralement les centres d'intérêts de l'utilisateur.

La représentation ensembliste du profil utilisateur apporte l'avantage de lasimplicité de mise en œuvre. Cependant, même si les modèles de représentationensembliste permettent de traduire une multiplicité des centres d'intérêts del'utilisateur, cette représentation manque de structuration, de cohérence, desniveaux de généralité/spécificité et des relations de corrélation entre les diverscentres d'intérêts de l'utilisateur [Saoud 2017].

3.6.3. Représentation connexionniste

La représentation connexionniste du profil utilisateur consiste à représenter les centres d'intérêts de l'utilisateur par un réseau de nœuds pondérés dont chaque nœud représente un concept traduisant un centre d'intérêt de l'utilisateur. Cette représentation permet de résoudre les failles de la représentation ensembliste par la mise en place des relations de corrélation sémantiques entre les centres d'intérêts du profil. En effet, la richesse sémantique dans cette représentation permet de résoudre le problème de la polysémie des termes inhérents à la représentation ensembliste, l'incohérence possible entre les centres d'intérêts et l'identification d'un profil adéquat au sujet de la requête via les relations sémantiques.

Cependant, cette représentation présente certaines limites. En fait, la source de données du réseau sémantique représentant le modèle utilisateur est l'historique de recherche de l'utilisateur, qui est généralement limité et nécessite l'utilisation de sources sémantiques externes telles que les ontologies [Daoud 2009].

3.6.4. Représentation multidimensionnelle

Dans cette représentation, le profil utilisateur est représenté par un ensemble de catégories (dimensions), où chaque catégorie contient des informations spécifiques sur l'utilisateur.

Les dimensions proposées se diffèrent d'une approche à l'autre. En effet, dans [Amato et Straccia 1999] est développé un service de bibliothèque numérique qui exploite un profil utilisateur multidimensionnel. Ce profil contient cinq catégories : la catégorie de données personnelles, la catégorie de données de la source, la catégorie de données de livraison, la catégorie de données de comportement et la catégorie de données de sécurité. Tandis que dans [Kostadinov et al 2008] est proposé un autre modèle de représentation multidimensionnelle de profil utilisateur contenant cinq autres dimensions : Le domaine

d'intérêt, les données personnelles de l'utilisateur, les données de livraison, les données de sécurité et la qualité

3.6.5. Représentation conceptuelle

Se basent sur l'exploitation des ontologies de domaines ou des hiérarchies de concepts préalablement définies. L'approche de représentation conceptuelle consiste tout d'abord à spécifier les niveaux des concepts de l'ontologie à considérer, et ensuite appliquer le procédé de déploiement des données dans des techniques de pondération de ces concepts. A la fin, le contexte utilisateur sera présenté par un réseau de nœuds conceptuels reliés entre eux en respectant la topologie des liens définis dans les hiérarchies ou les ontologies utilisées.

Les caractéristiques de l'utilisateur sont représentées sous forme de concepts. Les relations sémantiques entre les centres d'intérêts de l'utilisateur peuvent être obtenues par différentes techniques :

- Par l'utilisation des documents qui conservent le contexte documentaire des mots (centres d'intérêts de l'utilisateur), afin de construire des relations inter-documents entre ces mots [Billsus et Pazzani 1999].

- Par la construction de contextes, en associant pour chaque mot un ensemble de mots qui sont proches sémantiquement de lui [Marinilli et al 1999].

- Par la représentation des préférences et des centres d'intérêts d'utilisateur sous forme d'ontologie [Blanco-Fernández et al 2008 ; Daoud et al 2008].

3.7. Evolution de profil utilisateur

L'évolution des profils désigne leur adaptation à la variation des centres d'intérêt des utilisateurs qu'ils décrivent, et par conséquent, de leurs besoins en information au cours du temps. La phase d'évolution ne prend un sens que lorsque le profil a une structure pérenne, ce qui permet de distinguer les besoins à court terme, construits à partir de la session d'interaction courante, des besoins à long terme qui sont une réelle représentation des centres d'intérêt persistants de l'utilisateur. Peu de travaux ont exploré le problème de l'évolution du profil de l'utilisateur sous l'angle de la dimension temporelle (court terme, long terme).

Dans la plupart des systèmes d'accès personnalisé, l'évolution du profil consiste à adapter son contenu aux variations des besoins utilisateur en information :

- Dans le cas de systèmes à représentation ensembliste du profil, son évolution est exprimée par l'ajout de nouveaux vecteurs de termes extraits des documents correspondants aux centres d'intérêt détecté de l'utilisateur [Billsus et Al 1999].

- Dans le cas d'une représentation hiérarchique, le système fait évoluer le contenu du profil en associant de nouveaux documents collectés aux classes similaires appropriées. L'adaptation de la structure du profil aux nouvelles classes s'effectue en mettant à jour les relations entre ces classes [Moukas1997]. Elle traduit dans [Chen et al, 2002] la notion de

cycle de vie d'un centre d'intérêt, où on associe aux classes du profil utilisateur une valeur d'énergie, traduisant le degré d'importance d'un centre d'intérêt par rapport à un autre. Basée sur les valeurs d'énergie des catégories, la structure de profil peut être modulée pendant que les centres d'intérêt des utilisateurs changent.

- Dans [Zemirli 2008] l'évolution de profil consiste en une évolution de la dimension centre d'intérêt basée sur une mesure de corrélation des rangs qui évalue le degré de changement entre contextes d'usage associés à des périodes successives.

4. Les dimensions du contexte de recherche

De nombreux travaux en recherche d'information [Borlund 2003; Taylor et al. 2007; Saracevic 2007], ont prouvé que le concept de pertinence a un caractère multidimensionnel, et ont identifié plusieurs critères, au-delà de la dimension thématique, qui peuvent affecter la perception de pertinence de l'utilisateur. Ce qui a poussé vers une direction de recherche qui consiste à considérer, à la fois, tous les critères ayant un impact sur la notion de pertinence et pouvant influencer les jugements de pertinence des utilisateurs, afin d'améliorer la précision des systèmes de recherche.

Pour améliorer la représentativité des profils et leur capacité à décrire les intérêts des utilisateurs et d'inférer leurs besoins en information, les approches de la recherche d'information personnalisée exploitent, comme capteurs de pertinence utilisateur :

- En premier lieu, les données thématiques issues de l'historique de recherche pertinent de l'utilisateur. En effet, le contenu (thèmes, titres, termes, mots-clés, …etc.) des documents jugés pertinents par l'utilisateur est une source large et riche d'information qui peut mieux aider à identifier les centres d'intérêts de l'utilisateur. La dimension de pertinence thématique ou topicale qui traduit la proximité des sujets des requêtes et de documents est la plus couramment utilisée, et plusieurs travaux ont souligné que ce critère thématique est la dimension principale de pertinence et que tous les autres critères en sont dépendants [Moulahi 2015].
- En second lieu, d'autres capteurs supplémentaires de pertinence liés aux différentes dimensions contextuelles de recherche sont introduits pour mieux cerner la pertinence utilisateur. Un nombre considérable d'études ont souligné l'aspect multidimensionnel de la pertinence utilisateur.On cite la dimension sociale, localisable, temporelle, et même sémantique.

Nous présentons dans ce qui suit, les différentes dimensions contextuelles des informations représentant des capteurs de pertinence utilisateur, que le profil d'utilisateur doit les identifier et les exploiter pour l'amélioration de la performance des résultats de recherche. Il est à noter, comme c'est affirmé dans [Buidghaghen 2011], qu'il est difficile de trouver un système de recherche personnalisé qui gère toutes les dimensions contextuelles en même temps. Les dimensions à considérer lors de la prise en compte du contexte peuvent être plus ou moins pertinentes selon l'application de recherche.

4.1. Dimension sociale

Dans le cadre de la recherche d'information personnalisée, les informations sociales sont utilisées pour construire les profils des utilisateurs. Deux informations sociales principales sont utilisées :

- ***Les annotations sociales (Tags)***
  Les tags employés par les utilisateurs pour annoter les documents sont souvent utilisés pour déterminer leurs centres d'intérêts. Les annotations sociales sont très significatifs pour caractériser l'utilisateur de fait qu'ils expriment un avis, un jugement ou une impression sur les contenus des documents [Kamer et al. 2014 ; Ould Amer 2020].
  Plusieurs travaux ont été proposés pour estimer le poids de chaque tag, où ce poids reflète son importance pour l'utilisateur. Certaines approches proposent une représentation vectorielle des tags de l'utilisateur. Le poids de chaque tag peut suivre un schéma de pondération basé sur la fréquence d'utilisation de ce tag par l'utilisateur (TF) [Noll et Meinel 2007], ou par un TF-IDF [Xu et al. 2008], ou une combinaison du modèle TF-IUF et du modèle BM25 [Vallet et al. 2010].
  Selon [Heymann et al. 2008], il est constaté que 50% des tags utilisés pour annoter les documents sont des termes du document. Ceci démontre que certains utilisateurs ont tendance à employer des termes du document comme tags pour les annoter car de leur point de vue ces termes décrivent bien le sujet du document. Une autre étude [Bishoff et al. 2008], a montré que les tags des utilisateurs sont corrélés avec leurs motivations d'annotation. Un utilisateur peut annoter un document dans le but de le retrouver par la suite (par exemple tag : *à lire*), ou pour l'envoyer à quelqu'un (par exemple tag : *à partager*), …etc.
- ***Les relations sociales entre utilisateurs du système de recherche***
  L'exploitation des relations sociales entre utilisateurs du système de recherche permet d'enrichir le profil de l'utilisateur, où la pertinence prédite d'un document donné peut être améliorée en explorant les actions du voisinage de l'utilisateur. La notion d'importance sociale des utilisateurs est aussi introduite, et elle suppose qu'un document pertinent est produit par un auteur important d'où l'importance scientifique des ressources informationnelles peut être dérivée à partir de l'importance sociale des auteurs associés [Bouhini 2014 ; Benjabeur 2010].
  Les liens d'amitié avec les autres utilisateurs de système de recherche peuvent aider à deviner le besoin en information de l'utilisateur, de fait que les besoins cognitives des individus peuvent être influencés par ce qui intéresse et attire l'attention de leurs communautés proches (amis, voisins, collègues,…) [Badache et al. 2019]. En effet, certaines études selon [Smyth 2011] affirment que (66 %) de nos recherches sont similaires à celles de nos collègues.

Plusieurs approches de la RI personnalisée, ces dernières années, exploitent les annotations et les relations sociales pour la construction des profils des utilisateurs. Badache et al. [2019] proposent, dans un cadre de recherche d'information sociale en langue arabe sur Facebook, une approche qui sélectionne et exploite les meilleurs traces des utilisateurs

(polarité, partage, j'aime, haha) laissées sur des publications Facebook pour estimer leur importance sociale. L'objectif est de montrer comment ces signaux peuvent jouer un rôle vital dans l'amélioration de la recherche en langue arabe sur Facebook. Monteserin et Armentano [2019] proposent une approche de prédiction de liens entre les utilisateurs d'un réseau sociale. Ils définissent la notion d'influence sociale pouvant être exercée sur un utilisateur donné par les autres utilisateurs de réseau par tendance de l'utilisateur à accepter leurs suggestions et recommandations. Ils exploitent la transitivité pour prédire les liens indirects. Bouhini et al. [2016], proposent une approche de personnalisation de la recherche basée sur la combinaison des annotations sociales de l'utilisateur et les termes pertinents de son profil. Saoud et Kechid [2016] présentent, dans un cadre de recherche d'information distribuée, une approche de personnalisation combinant les annotations et les relations sociales de l'utilisateur pour la construction de son profil. Ould Amer et al. [2015] proposent une approche de recherche de conversations dans les tweets basée sur la combinaison de pertinence textuelle et sociale des tweets. Plusieurs facteurs sont considérés dans la pertinence sociale de la tweet comme les hashtags (#) et les URLs qu'elle contient, ainsi que l'expertise du blogueur qui la publié et son influence sociale représentée par le nombre de ses tweets favorisées ou retweetées par les autres blogueurs. Shafiq et al. [2015] ont proposé une nouvelle approche de personnalisation de la recherche sur le Web, qui extrait les préférences des utilisateurs en se basant sur les activités social des utilisateurs sur un réseau social. Les préférences de l'utilisateur sont exploitées pour re-classifier les résultats retourné par le moteur de recherche.

4.2. Dimension localisable

L'orientation vers l'introduction de la dimension localisation de recherche est motivée par la prolifération de la technologie mobile (PDAs, téléphones mobiles, tablettes, …) qui a rendu l'accès à une grande masse de documents sur le web, possible à toute place et à tout moment. En effet, vu les contraintes et particularités techniques des appareils mobiles (des difficultés de saisie de requêtes, zone d'affichage limitée, …), nous assistons à une pratique de recherche différente de celle des requêtes traditionnelles de bureau.

L'hypothèse de base de cette orientation est que la pertinence d'un document pour une requête donnée dépend de l'emplacement de l'utilisateur au moment de formulation de la requête. Des études [Buidghaghen 2009], montrent que 25 % des besoins informationnels des utilisateurs mobiles sont liés à la localisation. On peut imaginer, par exemple, si un utilisateur lance la requete « *echourouk* » en Algérie, fort possible qu'il cherche le journal algérien *Echourouk*, mais s'il l'utilisateur est en Egypte, la réponse serait mieux le journal *Echourouk* de l'Egypte. Un autre exemple, si un utilisateur en déplacement qui émet la requête "*restaurant pas cher*" s'intéresse plus aux restaurant proches de sa localisation que ceux qui sont dans une autre zone géographique. Donc le critère emplacement est très important dans ce contexte de recherche.

Plusieurs travaux en recherche d'information personnalisée exploitent l'information situationnelle ou de localisation (coordonnées GPS, nom de localisation, type de localisation, …etc.) et proposent de filtrer l'information sur la base de la proximité géographique pour améliorer les résultats de recherche d'information. Les auteurs dans

[Akermi et al. 2015], proposent une approche de recommandation personnalisée dans un environnement mobile guidée par un profil situationnel de l'utilisateur. Ce profil est basé sur la localisation et les préférences de l'utilisateur, et vise à identifier la bonne information à recommander en fonction de la situation dans laquelle se trouve l'utilisateur. O. Bouidghaghen [2011], propose une approche de personnalisation de la recherche basée sur la construction d'un profil situationnel d'utilisateur composé des différentes situations de recherche et les centres d'intérêts attachées, et elle propose un modèle de classification des requêtes pour identifier leurs sensibilités à la localisation. Dans [Bila 2008], les auteurs proposent un profil d'utilisateur mobile dans lequel les situations de l'utilisateur sont représentées par les lieux les plus visités par l'utilisateur, et les relations entre elles sont modélisées à base d'arbre. Les centres d'intérêt de l'utilisateur sont explicitement acquis à l'aide d'un questionnaire. Dans [Hattori et al. 2007], les auteurs traitent le problème de l'ambiguïté des requêtes en utilisant une méthode d'expansion basée sur la localisation. Leur méthode consiste à récupérer les coordonnées spatiales de l'utilisateur, les transformer en des mots contextuels (noms de places, d'activités liés à cette place) en utilisant un système d'information géographique et des techniques d'apprentissage des weblog, puis à pondérer ces mots contextuels sur la base de la comparaison de la probabilité globale du mot contextuel dans l'ensemble des documents du corpus cible et sa probabilité locale dans les documents retournés par la requête originale, en fin, la requête originale est étendue par les mots contextuels ayants les poids les plus élevés.

4.3. Dimension temporelle

Une autre dimension très importante pour la description des intérêts de l'utilisateur est la dimension temporelle ou évènementielle.La recherche d'information temporelle est un domaine de recherche dont l'objectif principal est l'amélioration des modèles classiques de la recherche d'information en exploitant les informations temporelles qui peuvent exister dans les requêtes et les documents. D'une façon générale, les modèles de recherche d'information combinent la notion de pertinence traditionnelle avec la dimension temporelle. En effet, dans certains cas le besoin réel d'information d'une requête utilisateur est incité par un événement important qui occupe l'opinion public sur le réseau social. Et ainsi, la prise en compte des évènements et des informations temporelles par l'extraction à partir des textes et la représentation dans le profil utilisateur peut bien aider à améliorer la pertinence des résultats de recherches [Moulahi 2015].

L'information temporelle peut concerner le temps de soumission des requêtes utilisateurs ou le temps de publication des documents ciblés, comme elle peut concerner les expressions temporelles ou celles relatives aux évènements dans les contenus des documents.

Dans [Vandenbussche et al. 2011] est proposée une approche recherche d'information qui combine les informations temporelles sur le document et celles de son contenu. Les auteurs de [Kanhabua et al. 2012] propose un système temporel-sémantique pour l'apprentissage de l'extraction d'évènement et d'entités nommées à partir de texte. Dans [Arikan et al. 2009] la requête utilisateur est considérée porteuse de deux sens thématique et temporel. Le travail présenté dans [Basil et al. 2014] propose un système pour l'extraction et

l'indexation d'évènements et la récupération des sujets liés à ces évènements à partir des textes. Dans [Panayiotou et al. 06], les auteurs exploitent l'importance du temps et de l'expérience (au travail, en vacances, etc.) dans la personnalisation d'un portail de recherche de services web pour un utilisateur mobile. Ils proposent de construire un profil dynamique où les centres d'intérêt sont pondérés selon des zones temporelles apprises par l'étude de la routine journalière de l'utilisateur et ses activités dans chaque zone.

4.4. Dimension sémantique

Plusieurs travaux de la recherche d'information personnalisée représentent le profil utilisateur à l'aide de l'ensemble de termes ou de mots clés issus de l'historique de recherche pertinent [Uddin et al. 2012]. Néanmoins, un processus de recherche basé uniquement sur la syntaxe (mots clés/termes), semble plus proche à une opération de reconnaissance de formes, qu'à une opération d'assimilation et de satisfaction d'un besoin cognitif de l'utilisateur exprimé par une requête souvent courte et ambigüe [Aimé et al. 2010]. Et sans l'exploitation des liens et des relations sémantiques entre les termes, la recherche d'information reste incapable de bénéficier de la richesse de la langue humaine pour une compréhension profonde de la requête, du besoin de l'utilisateur et du document cible [Fouad et al. 2012]. En effet, beaucoup de documents dont le contenu est sémantiquement similaire au sujet de la requête peuvent ne pas être sélectionnés, de fait qu'ils utilisent d'autres termes pour décrire différemment le même contenu informationnel.

La sémantique ne considère pas les documents comme un simple ensemble de termes sans liens entre eux, mais plutôt elle tient compte de la signification véhiculée par ces termes. Les approches de recherche d'information personnalisée orientée sémantique, s'appuient sur l'exploitation des ressources sémantiques externes tels que les ontologies, les thesaurus, les dictionnaires…etc. pour extraire les liens sémantiques entre les termes [Zargayouna et al. 2015].

Plusieurs travaux en recherche d'information personnalisée s'orientent, ces dernières années, vers l'introduction de la sémantique dans le profil d'utilisateur pour améliorer l'exactitude et la pertinence des résultats de recherche en mieux comprenant les intentions de l'utilisateur. Mohammed et Abdelmoty [2017] proposent un profil spatial sémantique dans un réseau social basé sur les localisations. Ce profil est construit par la combinaison des informations concernant les différentes localisations visitées par l'utilisateur, et les annotations sociales qu'il a utilisées pour annoter ces localisations ou pour annoter les activités qu'il a effectuées dans ces localisations. Les auteurs utilisent la base de localisation « Foursquare » pour extraire les informations sémantiques des localisations (nom, catégorie, région,…), et la base lexical de l'anglais « WordNet » pour filtrer et catégoriser les annotations de l'utilisateur. T. H. Duong et al. [2013] présentent une approche de personnalisation sémantique de la recherche basée sur l'ontologie ODP (open directory project). Les concepts de l'ontologie sont utilisés pour l'annotation des documents pertinents de l'historique de l'utilisateur, afin de créer un profil sémantique. Ces concepts sont utilisés aussi dans l'indexation sémantique des documents pour créer un espace de recherche sémantique. Daoud et al. [2011], dans leur approche de personnalisation sémantique proposent de modéliser le profil d'utilisateur et les documents

sous forme d'un graphe de concepts. Ces concepts sont issus de l'ontologie ODP. Ils proposent aussi un modèle de ré-appariement (doc/profil) basé sur une mesure de distance sémantique qui combine le maximum et le minimum sous graphe de concept commun. Dans [Schenkel et al. 2008], en plus de la dimension sociale sur laquelle les auteurs se basent pour construire un contexte social de l'utilisateur, ils proposent de prendre en compte une dimension sémantique pour sélectionner des termes. Cela se fait sur la base d'une similarité sémantique entre ces termes et ceux de la requête de l'utilisateur.

5. Intégration de profil utilisateur dans le processus de recherche

Après que le profilest formé, ce dernier doit êtreintégré au système de RI afin d'inférer les nouveau besoins en information de l'utilisateur. Il existe principalement deux phase pour intégrer le profil utilisateur : (1) Reformulation de la requête (2) reclassement des résultats de la recherche.

5.1. Reformulation de requête

Un grand nombre de méthodes de recherche d'information a été développé afin d'aider les utilisateurs à reformuler ou à affiner leur requête. Ces méthodes sont classées comme automatique ou interactives[rwthven 2003].

La reformulation de requête est un processus qui améliore automatiquement la requête initiale de l'utilisateur en ajoutant des termes du profil de l'utilisateur et cela pour identifier les meilleurs termes pour enrichir la requête de l'utilisateur.

L'approche dans [tamine et al. 2008] intègre le profil utilisateur dans la fonction d'appariement du model de recherche bayésien. La valeur de pertinence d'un document vis-à-vis d'une requête n'est plus fonction de la requête seule mais en plus du centre d'intérêt de l'utilisateur qui la soumise.

Dans Xie et al,[xie et al 2012]la requête initiale formulée par un utilisateur u est alors représentée par un ensemble de termes pondérés, comme montré dans l'équation suivantes

$$q_u = (t_1^{qu} : w_{qu,t1}, \ldots, t_j^{qu} : w_{qu,tj} \ldots, t_m^{qu} : w_{qu,tm})$$

Où,t_j^{qu} représente un terme t dans la requête q_u de l'utilisateur u et $w_{qu,t}$est le poids du terme t dans la requête q_u de l'utilisateur$u$ tel que $w_{qu,t}$= 1 La requête étendue composée des termes initiaux de la requête, combinés aux termes en provenance du profil social de l'utilisateur, est donnée par la formule

$$q'_u = (t_1^{q'} : w'_{qu,t1}, \ldots, t_j^{q'} : w'_{qu,tj}, \text{tj} \ldots, t_n^{q'} : w'_{qu,tn}) q'_u$$

Où, le poids d'un terme dans la requête étendue par les termes des annotations sociales $w'_{qu,t}$ est donné par une combinaison linéaire des termes initiaux de la requête de l'utilisateur et les termes du profil social de cet utilisateur :

$$w'_{qu,t} = \alpha w_{qu,t} + (1 - \alpha) w_{u,t}$$

Où, $w_{u,t}$ est le poids du terme t dans le profil de l'utilisateur u.

La requête ainsi améliorée par les données du profil est nommée « requête profilée ».

5.2. Reclassement de résultats

Les moteurs de recherche utilisent des algorithmes de reclassement qui leur permettent de classer des résultats suivant leur pertinence à une requête. A l'aide de données constituées de pair requête-document pour lesquelles la pertinence est connue. Un grand nombre d'approches et d'algorithmes ont été développés pour le but de reclasser les résultats.

Les approches basées sur le réordonnancement des résultats de recherche [Challam et al 2007 ; Sieg et al 2007]consistentsouvent à combiner le score original du document et son score de similarité avec le profil d'utilisateur. Des variantes des approches de réordonnancement des résultats consiste en une catégorisation personnalisée [Ma et al. 2007]basée sur la classification des résultats de recherche dans des catégories représentatives des centres d'intérêts du profil utilisateur. La personnalisation dans [liu et al. 2004] consiste à créer plusieurs listes de résultats associées aux catégories associées a la requête, ensuite à les combiner selon une méthodes de réordonnancement par vote majoritaire.

6. L'évaluation des systèmes de recherche d'information personnalisés

6.1. Les limites de l'évaluation classique

L'évaluation des SRI est l'un des piliers de l'évolution de la recherche d'information. La qualité de l'évaluation est d'une importance capitale puisqu'elle permet de discriminer les différents modèles. L'évaluation a pour but de mesurer la capacité des SRI à satisfaire les besoins d'information des utilisateurs.

Généralement, un système de RI classique est évalué à travers d'une collection de test comprenant 3 éléments essentiels qui sont [Ouled Amer 2020] :

- Un ensemble de requêtes qui représentent les besoins en information des utilisateurs. Cet ensemble est fourni afin d'évaluer pour chaque requête la capacité du système à retrouver les documents pertinents.
- Un ensemble de documents qui contiennent les informations pertinentes qui répondent à l'ensemble des requêtes des utilisateurs.
- Les jugements de pertinence qui permettent d'identifier les documents pertinents pour chaque requête. Ces jugements sont fournis généralement par des utilisateurs.

Avec toutes les performances de l'évaluation classique qui existe dans les différentes compagnes d'évaluation comme TREC, CLEF ou autres, mais cette stratégie d'évaluation présente une limite en prenant compte de l'utilisateur. Les tests réalisés ne donnent pas importance ni au contexte dans lequel se fait la recherche, ni à la perception de la pertinence de l'utilisateur dans ce contexte, ni à la diversité des centres d'intérêts des utilisateurs. Par conséquent il est difficile d'évaluer les collections de test

qui traduisent la subjectivité de la notion de pertinence et des centres d'intérêts des utilisateurs.

6.2. La problématique de l'évaluation personnalisée

D'après [Tamine et Daoud 2018 ; Toms et al. 2013], dans le domaine de personnalisation de la recherche d'information, la pertinence à améliorer est trop subjective, vue qu'elle repose particulièrement sur l'utilisateur qui est le seul qui peut juger cette pertinence selon son point de vue. C'est pour cela que l'évaluation d'une approche de recherche d'information est problématique, due à l'absence d'une collection de tests avec des requêtes et des pré-jugements faites par les utilisateurs concernés par l'étude eux-mêmes.

L'évaluation d'un système de recherche d'information personnalisée représentent un véritable défi, car ça nécessite de fournir les composantes principales suivantes [Bouhini 2016] :

- Une collection des informations sur les utilisateurs, par convention appelée profil d'utilisateur, servant à l'organisation, la sauvegarde et l'exploitation des données sur l'utilisateur et son historique de recherches et d'interactions avec le système.
- Une Dataset (collection de documents + ensemble de requêtes), pour laquelle doit être associé un ensemble de pré-jugements de pertinence fait par les mêmes utilisateurs concernés. Les pré-jugements des utilisateurs servent pour référence d'évaluation des performances des nouvelles approches de personnalisation de la recherche.

La complexité est qu'il y'a pas decollections de tests standards pour l'évaluationde l'efficacité des systèmes de recherche d'information personnalisée. Fournir de telles collections est très complexe par le fait de la difficulté de la mise en place d'un système qui permet de collecter les informations des utilisateurs, leurs requêtes et ensuite les jugements de pertinence pour tous les documents pour chaque requête pour chaque utilisateur [Ouldamer 2020].

6.3. Protocole de l'évaluation personnalisée

Vue l'absence d'un cadre d'évaluation standard ou consensuel dédié à la recherche d'information personnalisée, les approches de ce domaine ont eu recours à développer leurs propres cadres d'évaluation.

Dans ce qui suit, nous citons quelques exemples de solutions:

(i) Collection des informations sur l'utilisateur

Pour la collection des données d'utilisateurs pour la construction des profils, Bouhini (2016) et Saoud and Kechid (2016) ont utilisé l'historique de *Bookmaring* du site social *Delicieuse.* Daoud et al. (2011) ont utilisé les fichiers log de recherche sur le site web commercial *Exalead.* Mohammed et Abdelmoty (2017) ont utilisé l'historique de chek-in faites par les utilisateurs de la base de données géographique *Foursquare* pour une période donnée. T.H. Duong et al. (2013), ont utilisé *Google scholar citation* et *Facebook* pour la

collection des données d'utilisateurs. Buidguaguen (2011) propose un algorithme basé sur les interactions utilisateur-requête pour la simulation du profil d'utilisateur. Tandis que Bouramoul et al. (2011) interrogent l'utilisateur directement par un ensemble de questions pour définir son profil.

(ii) Dataset

Concernant les Dataset utilisées pour tester les performances des approches proposées, le problème est l'absence des pré-jugements faits pas les utilisateurs qui permettent de mesurer la pertinence des résultats de recherche. Pour cette raison, les approches de RIP proposent des solutions modifiées pour résoudre ce problème : Daoud et al. (2011), utilisent comme *Dataset* les 50 premiers documents retournés par le moteur de recherche*Yahoo* sur chacune des 50 requêtes envoyées, tant que Buidguaguen (2011) a retenu les 150 premiers documents pour chacune de 25 requêtes envoyées.Saoud et Kechid (2016) qui s'intéressent à la recherche distribuée, utilisent 10 moteurs de recherche différents, et ont retenu comme *Dataset* les tops 20 documents retournés par chaque moteur de recherche sur 90 requêtes envoyées. Du même, Bouramoul et al. (2011), considèrent comme *Dataset*, les tops 20 documents retournés par les 3 moteurs de recherche *Google, Yahoo* et *Bing* pour chacune des 30 requêtes envoyées.

(iii) Evaluation des performances

Comme ce genre de *Dataset* manque de pré-jugements de pertinence des utilisateurs sur lesquels les données sont collectées et les profils sont construits, les approches de RIP les remplacent par des post-jugements faits par des individus dits « *Participants* ». Cinq participants sont invités dans [Daoud et al.2011] et [Duang et al.2013] pour évaluer la pertinence des résultats retenus. Et respectivement, 6,15 et 24 participants sont invités dans [Buidguaguen2011], [Saoud et Kechid2016] et [Bouramoul et al.2011]. Selon les post-jugements de pertinence faits par ces participants, les améliorations des approches proposées sont évaluées.

6.4. Les mesures de l'évaluation des systèmes personnalisée

- *La mesure de précision P@K*

La performance des systèmes est généralement évaluée à partir de la précision à la position k P@k, de la précision moyenne AP et de la moyenne de la précision moyenne sur l'ensemble des requêtes MAP, définies ci-dessous.

$$P@k = \frac{Nombre\ de\ documents\ pertinents\ sélectionnés\ jusqu'au\ rang\ k}{Nombre\ totalde\ documents\ pertinents\ jusqu'au\ rang\ k}$$

- *F-Mesure*

Est une mesure qui estime une moyenne harmonique pondérée de la précision et du rappel et est décrite par la formule suivante :

$$F - mesure = \frac{2 * Précision * Rappel}{Précision + Rappel}$$

- *le NDCG (Normalized Discounted Cumulative Gain)*

Elle permet d'évaluer la capacité des SRI à renvoyant les documents pertinents en haut de la liste de résultats. Elle est définie de la façon suivante:

$$NDCG@K = \frac{DCG}{IDCG@K} \text{Où} DCG@K = \sum_{i=1}^{k} \frac{2^{rel(k)}-1}{log_2 i}$$

EtIDCG@k est la valeur maximale possible de DCG@k.

- *La MAP (Mean Average Precision)*

Est une mesure qui estime la moyenne des précisions moyennes sur l'ensemble des requêtes permettant ainsi de mesurer la performance global d'un système, et est décrite par la formule suivante :

$$MAP = \frac{1}{|Q|} \sum_{q \in Q} \frac{1}{r} \sum_{r \in R} précision\ (q)@\ R$$

- *Le P-Gain*

Mesure le gain personnalisé d'un système par rapport à un autre système. Les taux sont calculés aux niveaux des performances au niveau de la requête en relevant les différences en valeurs de précision moyenne AP (AveragePrecision) obtenues entre deux systèmes. Cette mesure est estimée comme suit :

$$P_Gain = \frac{\#Q_+ - \#Q_-}{\#Q_+ + \#Q_-}$$

Où$\#Q_+$ représente le nombre de requêtes améliorées, $\#Q_$représente le nombre de requêtes détériorées par rapport à un système.

7. *Conclusion*

La personnalisation de l'information immerge comme une approche capitale dans le développement des futurs systèmes de recherche. Son but est de nous fournir une information pertinente, qui correspond exactement au besoin de l'utilisateur.

Dans ce chapitre nous avons définit la recherche d'information personnalisé RIP. Ensuite nous avons présenté la modélisation de l'utilisateur qui a pour objectif d'identifier, collecter et organiser les informationssur l'utilisateur et caractérise sa pertinence. De plus, nous avons présenté la structure d'un profil d'utilisateur qui représente la concrétisation de

la modélisation. Nous avons présenté les sources d'informations, les méthodes d'acquisition et celles de représentation des données du profil, ainsi que les différentes dimensions couvertes par ces données à savoir la dimension thématique, sociale, mobile ou même sémantique. Ensuite, nous avons entamé le sujet d'intégration du profil d'utilisateur dans le processus de recherche, et pour terminer nous avons donnés un aperçu sur l'évolution d'un profil utilisateur dans un SRIP.

Chapitre 03

Exploitation des liens d'amitiés dans un site social pour l'amélioration du résultat de recherche

Chapitre3 :Exploitation des liens d'amitiés dans un site social pour l'amélioration du résultat de recherche

1. *Introduction*

Personnaliser l'information signifie de s'adapter aux buts, préférences et capacités de l'utilisateur. La personnalisation a pour but de facilites l'expression de besoin utilisateur et de rechercher des informations sur un sujet en écartant l'information non-pertinente et réduire l'espèce de recherche et de rendre cette information sélectionnée intelligible à l'usager et exploitable. Et ainsi, l'objectif des systèmes de recherche d'information est d'inférer le réel besoin en information de l'utilisateur caché derrière sa requête, et de lui retourner par la suite les résultats les mieux correspondants à son besoin et non seulement à sa requête [Daoud 2011]. Plusieurs études montrent que les requêtes d'utilisateurs sont souvent courtes et ambigües et qu'elles n'expriment pas clairement le besoin en information de l'utilisateur. De ce fait, la requête d'utilisateur a besoin d'être enrichie par l'injection de certains autres termes en relation avec son sujet pour qu'elle puisse mieux exprimer le besoin en information de l'utilisateur[Raza et al. 2018 ; Bendella and Quafafou 19].Pour réaliser cet objectif, les systèmes de recherche d'information personnalisée utilisent une structure, par convention appelée profil d'utilisateur [Nazim Uddin 2012], permettant une description détaillée de l'utilisateur et ses intérêts. Le profil de l'utilisateur est intégré par la suite dans le processus de recherche pour une meilleure adaptation des résultats de recherche aux intentions de l'utilisateur.

L'hypothèse de base des approches de RIP est que les nouvelles recherches des utilisateurs sont souvent en relation avec leurs recherches précédentes. Et par conséquent, le besoin en information de l'utilisateur peut être inféré en investiguant son historique de recherche. De ce fait, les profils d'utilisateurs ne se limitent pas aux données statiques explicitement exprimées par l'utilisateur telles que l'identité, les domaines d'intérêts, les préférences…etc. mais ils donnent plus d'importance aux données dynamiques qui sont implicitement extraites à partir de l'historique de recherche pertinent de l'utilisateur. [Bouhini 2016 ; Saoud et Kechid 2016].

Pour améliorer la représentativité des profils et leur capacité à décrire les intérêts des utilisateurs et d'inférer leurs besoins en information, certaines approches exploitent, en plus des données thématiques (documents et termes), les informations sociales de l'utilisateur comme capteur supplémentaire de pertinence [Saoud et Kechid 2016 ; Benjabeur et al. 2011]. Deux informations sociales principales sont utilisées : (i) les tags utilisés par l'utilisateur dans les annotations des documents pertinents, qui sont très significatifs pour caractériser l'utilisateur de fait qu'ils expriment un avis, un jugement ou une impression sur les documents [Kumar et al. 2014], et (ii) les liens d'amitié avec les autres utilisateurs de système de recherche, de fait que les besoins cognitives des individus peuvent être influencés par ce qui intéresse et attire l'attention de leurs communautés proches (amis, voisins, collègues,…) [Badach et al. 2019]. En effet, certaines études selon [Smyth 2011] affirment que 66 % de nos recherches sont similaires à celles de nos collègues.

Dans ce chapitre, nous proposons une approche de personnalisation de la recherche d'information basée sur la construction d'un profilsociald'utilisateur. Dans ce profil nous combinons les capteurs de pertinence provenant de :

- Le contexte Thématique, représenté par les termes pertinents extraits à partir des documents pertinents des recherches précédentes de l'utilisateur.
- Le contexte Social interne, représenté par les tags les plus utilisés par l'utilisateur pour annoter les documents pertinents.
- Le contexte social externe, représenté par les liens d'amitié les plus importants reliant l'utilisateur avec les autres utilisateurs de système de recherche.

Notre objectif dans cette approche, est d'exploiter les informations thématiques et sociales fournies par le profil d'utilisateur pour reformuler les requêtes initiales de l'utilisateur dans le but de les désambigüiser et les aider à mieux exprimer le réel besoin en information de l'utilisateur, et par conséquent pour améliorer la pertinence des résultats de recherche.

Après cette introduction, le reste de document est organisé comme suit : la section 2 nous présentons notre profile d'utilisateur proposée, Le processus de recherche est présenté dans la section 3, suivi par l'évaluation de l'approche dans la section 4. Et nous terminons dans la section 5 par une conclusion.

2. *Profil utilisateur*

On appelle profil utilisateur toute structure qui permet de modéliser et de stocker les données caractérisant l'utilisateur. Ces données représentent les centres d'intérêts, les préférences et les besoins en informations de l'utilisateur ou un groupe d'utilisateurs[BK 05], [ZTB 05].

Les profils d'utilisateurs que nous proposons dans cette approche visent d'abord à mettre en évidence certains capteurs de pertinence à partir de l'historique de recherche de l'utilisateur et de l'historique des relations sociales avec d'autres utilisateurs du système de recherche. Deuxièmement, les capteurs associés sont évalués dans la vue pseudo-sémantique basée sur les liens de cooccurrence.

Le profile utilisateur proposé se compose de 4 composantes comme l'indique figure 1:

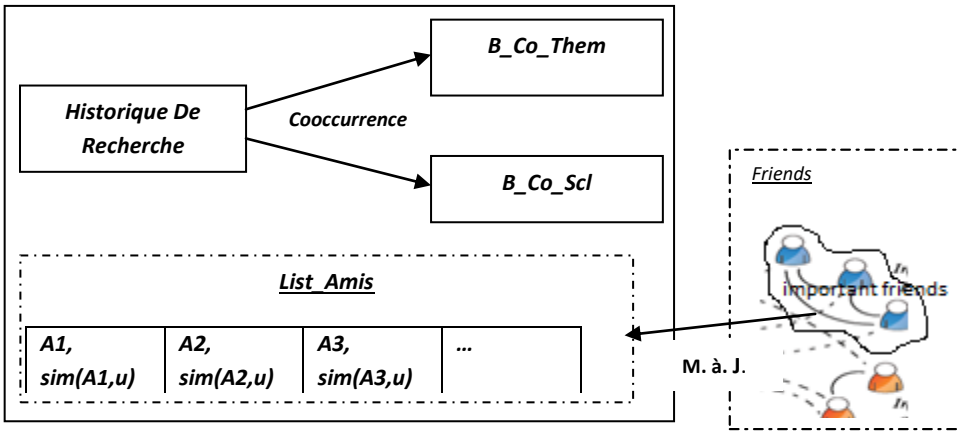


Figure 3.1. Profil d'utilisateur

2.1.L'historique de recherche pertinent

Cette composante (figure2) sert pour sauvegarder les traces des activités de recherche les plus significatives pour l'utilisateur. Elle se compose des 3 sous-ensembles suivants :

Historique de Recherche Pertinent
Documents Pertinents :   D1, D2, D3,…
Termes Pertinents : V_t= {(t1, wt1), (t2, wt2),…}
Tags Pertinents V_g= {(g1, wg1), (g2, wg2),…}

Figure 3.2. Historique de Recherche Pertinent de l'Utilisateur

2.1.1. Les documents pertinents

C'est l'ensemble des documents jugés pertinents par l'utilisateur (longuement visités, trop cliqués, sauvegardés, imprimés, partagés,…).

2.1.2. Les termes pertinents

Il s'agit d'une collection des termes les plus pertinents extraits de documents connexes. Ces termes représentent l'orientation thématique de l'utilisateur. La pertinence d'un terme est représentée par un poids qui représente l'importance du terme dans l'historique de recherche. Le poids du terme t, noté w(t), est calculé à l'aide de la formule tf*idf pour le modèle vectoriel [Salton et Buckley 1988]:

$$W(t)= (0.5 + 0.5 * tf/Max\ tf) * log\ (N/n) \quad (1)$$

Où :

- *tf* : fréquence totale de terme *t* dans les documents de l'historique.
- *Max tf :* la plus grande fréquence des termes dans les documents de l'historique.
- *N :* nombre de documents de l'historique.
- *n :* nombre de documents contenant le terme *t* de l'historique.

2.1.3. Les tags pertinents

Il s'agit d'un ensemble de tags que les utilisateurs utilisent le plus souvent pour annoter des documents connexes. Les balises représentent des capteurs de pertinence factuelle supplémentaires fournis par les utilisateurs eux-mêmes pour exprimer des opinions, des sentiments, des impressions, etc. Les étiquettes pondèrent également les historiques

d'annotation sociale de la même manière que les termes. Par conséquent, le poids de l'étiquette g est calculé comme suit :

$$W(g)= (0.5 + 0.5 * gf/Max\ gf) * log\ (N/n) \quad (2)$$

Où :

- *gf* : fréquence totale de tag *g* dans les documents de l'historique.
- *Max* gf : la plus grande fréquence totale des tags dans les documents de l'historique.
- *N* : nombre de documents dans l'historique.
- *n* : nombre de documents annotés par le tag *g* dans l'historique.

2.2. La base de liens de cooccurrence *terme-terme*

Comme mentionné ci-dessus, la collection de termes associés dans un centre d'intérêt représente une utilisation de premier niveau de l'historique de recherche pour décrire les utilisateurs et leurs intérêts. Le deuxième niveau d'utilisation des termes liés aux centres d'intérêt consiste à identifier et évaluer leurs relations de cooccurrence. Dans certains travaux de recherche d'information, cette relation est considérée comme une sémantique. En fait, les auteurs de [Mousset et al. 2019] soutiennent que la cooccurrence des termes lors de la description de différents objets textuels spatiaux est le lien sémantique entre ces objets. [Bernardo et al. 2013] ont proposé une méthode pour détecter les liens entre les ressources du réseau basée sur la cooccurrence d'entités dans les ressources.

Nous appliquons une pseudo-sémantique basée sur les relations de cooccurrence entre les termes pour rechercher des contextes de personnalisation. Nous souhaitons mettre en évidence les occurrences fréquentes de paires de termes associés dans de nombreux documents associés dans l'historique de recherche de l'utilisateur. Pour nous, l'occurrence fréquente des deux termes dans de nombreux documents suggère qu'ils coopèrent pour décrire le même sujet. Par exemple, l'apparition de deux termes « Java» et « Programmation » peut indiquer qu'il s'agit d'un sujet sur l'Informatique, tandis que l'apparition de « Java » et « ile » indique un sujet de tourisme. Et si « Java » figure dans une nouvelle requête utilisateur il serait facile de l'enrichir par l'ajout de terme ayant la relation la plus forte de cooccurrence avec « Java ».

La valorisation des relations de cooccurrence entre les termes pertinents est représentée par une similarité thématique, notée sim_them (ti, tj), qui exprime la liaison entre chaque terme pertinent ti avec les autres termes pertinents tj, et elle est calculée selon la formule suivante:

$$Sim_them\ (t_i,t_j) = |d(t_i,t_j)| / (|d(t_i)|+|d(t_j)-d(t_i,t_j)|) \quad (3)$$

Avec :

- $|d\ (ti,tj)\ |$: nombre de documents contenant ti et tj conjointement.
- $|d\ (ti)|$: nombre de documents contenant ti.
- $|d\ (tj)- d\ (ti,tj)\ |$: nombre de documents contenant tj et ne contenant pas ti.

Nous constatons que cette similarité de cooccurrence n'est pas bidirectionnelle, (sim_them (ti,tj)≠ sim_them (tj,ti)). En effet, il se peut qu'un terme 'x' apparaisse dans la majorité des documents contenant le terme 'y', tandis que 'y' n'apparait que dans une petite partie des documents contenant 'x'. Ceci traduit le fait qu'un terme peut être plus vaste qu'un autre.

La formule 3 favorise, pour un terme donné ti, les termes répondant aux 2 critères : (i) apparaissant dans le maximum de documents contenant $t_i$et (ii) apparaissant dans le minimum de documents ne contenant pas t_i.

La base de relations de cooccurrence des termes pertinents (B_Co_Them) comme l'affiche (figure 3), sert à sauvegarder les liens de similarité entre les termes pertinents de l'historique de profil utilisateur.

***B-Co-them*:**

Lien (term-term)

	t_1	t_2	…	t_n
t_1	–	Sim (t1,t2)	…	Sim (t1,tn)
t_2	Sim (t2,t1)	–	…	Sim (t2,tn)
..	…	….	–	…
t_n	Sim (tn,t1)	Sim(tn,t2)	….	–

Figure 3.3. Base de Cooccurrence Terme-Terme

2.3. La base de liens de cooccurrence termes-tags

Comme les tags d'utilisateurs peuvent être aussi utilisés pour l'enrichissement des requêtes, et dans le but de ne pas désorienter la recherche en d'hors de contexte actuel de l'utilisateur, un tag ne doit être injecté dans une requête que s'il possède une relation de cooccurrence très forte avec les termes de la requête.

Nous suivons la même logique précédente pour mettre en valeur les liens de cooccurrence termes-tags par une similarité sociale interne, notée *sim_scl* (t_i, g_j), qui exprime la liaison entre chaque terme pertinent t_i avec tous les aut0res tags pertinents g_j, et elle est calculée selon la formule suivante :

$$Sim_them\ (t_i,g_j) = |d(t_i,g_j)\ | \ / \ (|d(t_i)|+|d(g_j)-d(t_i,g_j)|) \qquad (4)$$

Avec :

- $|d(t_i,g_j)|$: nombre de documents contenant t_i et annotés par le tag g_j.
- $|d(t_i)|$: nombre de documents contenant t_i.
- $|d(g_j)-d(t_i, g_j)|$: nombre de documents annotés par g_j et ne contenant pas t_i.

La formule (4) proposée pour calculer la similarité sociale interne favorise, pour un terme donné t_i, les tags répondant aux 3 critères : (i) utilisés pour annoter le maximum de documents contenant t_i, (ii) absents dans le minimum de documents contenant t_i, et (iii) utilisés pour annoter le minimum de documents ne contenant pas t_i.

La base de relations de cooccurrence terme-tag (B_Co_Scl) comme l'affiche (figure 4), sert à sauvegarder les liens de similarité entre les termes pertinents et les tags fréquents de l'historique de profil utilisateur.

***B-co-Scl*:**

Lien (term-tag)

	g_1	g_2	…	g_k
t_1	Sim(t1,g1)	Sim (t1,g2)	…	Sim (t1,gk)
t_2	Sim (t2,g1)	…	…	…
..	…	….	… .	…
t_n	Sim (tn,g1)	Sim (tn,g2)	…	Sim (tn,gk)

Figure3.4. Base de Cooccurrence Terme-tag

2.4. Liste des amis les plus similaires (pertinence sociale externe)

Dans le cas de nouveaux ou de pauvre profil ainsi que le changement d'intérêt, le centre d'intérêt de l'utilisateur ne peut pas mener les nouvelles sessions de recherches. Pour cela les approches de la RIP demandent de l'aide auprès des amis les plus proches de l'utilisateur et qui partagent les mêmes intérêts avec lui. Le but est d'aider à désambigüiser la requête utilisateur et d'identifier le besoin en information qui peut être dérivé à partir des profils de ses amis.

Dans notre cas, nous dotons notre profil d'utilisateur par une liste dynamique des amis les plus similaires à l'utilisateur. Pour un utilisateur Ui, l'importance d'un ami (autre utilisateur) Uj est pondérée par une similarité, sim_ami(Ui, Uj), basée sur l'intersection des centres d'intérêts thématique et sociaux des deux utilisateurs.

La formule se présente comme suit :

$$sim_ami_{(u_i,u_j)} = \frac{\sum_{t \in CI_com(u_i,u_j)} w(t) + \sum_{g \in CI_com(u_i,u_j)} w(g)}{|\{v_t(u_i) \cup v_g(u_i)\}|} \qquad (5)$$

Tel que :

$CI_com(ui,uj)=\{v_t(u_i) \cup v_g(u_i)\} \cap \{ v_t(u_j) \cup v_g(u_j)\}$

Où :

- W(t) : le poids du terme t selon la formule 1.

- W(g) : le poids du tag g selon la formule 2.
- CI_com(ui, uj) : le centre d'intérêts commun entre les deux utilisateurs ui et uj.

Cette similarité est exprimée par la somme des poids de l'ensemble des termes et des tags communs dans les profils des deux utilisateurs par rapport au nombre total de termes et de tags dans le profil de l'utilisateur concerné u_i.

Périodiquement, la liste des amis les plus similaires à l'utilisateur List_Amis est mise à jour pour suivre la dynamicité et l'évolution de leurs intérêts. Seulement les amis dont la similarité dépasse la moyenne qui sont retenus.

2.5. Mesure de l'influence sociale sur l'utilisateur

La plus part des approches dans la RIP qui s'intéressent au sujet d'influence sociale d'un utilisateur donné par rapport aux utilisateurs d'un réseau social ou un groupe d'amis,expriment cette influence en se basant généralement sur le nombre de réactions positives (télécharger, regarder, aimer, partager, recommander, …) des utilisateurs avec des contenus ou des recommandations émanant de l'utilisateur.

Différemment de ces approches, nous nous intéressons à que l'utilisateur soit influencer par les autres utilisateurs. Autrement, nous voulons mesurer à quel point un utilisateur donné peut préférer ce que préfèrent les autres utilisateurs de système de recherche. Nous adoptons la définition de l'influence sociale citée dans [MonteserinArmentano 2019]« l'utilisateur A est influencé par l'utilisateur B s'il réagit positivement avec les suggestions et les recommandations provenant de l'utilisateur B ».

L'objectif est de mesurer à quel point il est utile d'utiliser les informations fournies par les profils des amis pour enrichir la requête initiale de l'utilisateur pour mieux exprimer son besoin en information.

Nous représentons cette influence sociale passive, ou la dépendance de l'utilisateur à ses amis proches, par un score noté DEP(U_i), exprimé par la proportion du nombre de documents pertinents trouvés dans les 5 Top résultats retournés pour les requêtes reformulées en utilisant les profils des amis, par rapport au nombre de documents pertinents trouvés dans les 5 Top résultats retournés pour les requêtes reformulées en utilisant le profil de l'utilisateur**.**

2.6. Mesure de l'indépendance sociale de l'utilisateur

Dans certains cas, l'utilisateur peut s'intéresser à des contenus peu intéressants à sa communauté proche. Ce que nous appelons la tendance d'indépendance sociale en matière de préférences et de perception de pertinence. Pour mesurer cette tendance d'indépendance sociale de l'utilisateur, nous proposons un score, noté INDEP(U_i), exprimé par la proportion du nombre de documents pertinents trouvés dans les 5 Top résultats retournés pour les requêtes reformulées en utilisant le profil de l'utilisateur, par rapport au nombre de documents pertinents trouvés dans les 5 Top résultats retournés pour les requêtes reformulées en utilisant les profils des amis.

Des valeurs élevées de ce score (INDEP (U_i)) indiquent qu'il n'est pas utile d'utiliser les profils de ses amis pour reformuler ses requêtes.

3. Processus de Recherche

Notre processus de recherche (figure 5) suit les étapes suivantes :

- Le système reçoit la requête utilisateur Q.
- Le système sélectionne la base de cooccurrence la plus correspondante à la requête parmi les deux bases (B_Co_Them et B_Co_Scl) de l'utilisateur, ainsi que celles de ses amis proches.
- Selon la base sélectionnée, la requête est reformulée pour produire une requête reformulée Q'.
- la requête reformulée Q' est envoyée au moteur de recherche classique.
- Le système observe les réactions et les comportements de l'utilisateur avec les résultats de recherche retournés pour mettre à jour son profil.

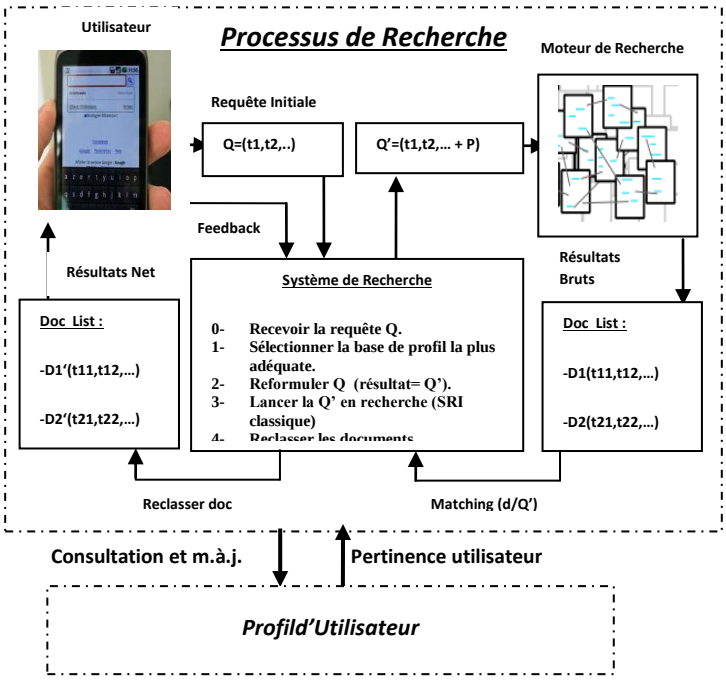


Figure3.5. Processus de Recherche.

Dans ce qui suit, nous présentons les différentes étapes du processus de recherche.

3.1. Sélection de la base la plus adéquate à la requête d'utilisateur

L'objectif de cette étape est de choisir parmi les différentes bases de liens de cooccurrence de profil de l'utilisateur ou des profils des amis proches, celle qui est plus proche au sujet de la requête et qui servira mieux à sa reformulation.

Pour pouvoir sélectionner parmi les deux bases de profil d'utilisateur (B_Co_Them et B_Co_Scl) et celles de ses amis, la base la plus correspondante à la requête utilisateur, notre système calcule des scores de similarité entre la requête et chacune de ces bases respectivement comme suit :

- A partir du profil de l'utilisateur :
  - $Sim\ (B_Co_Them,Q)=(INDEP(u))*log\ (\Sigma_{ti\in Q}\Sigma_{tj\in B_Co_Them} sim_them(t_i,t_j))$ (6)

  - $Sim\ (B_Co_Scl, Q)= (INDEP(u))*log(\Sigma_{ti\in Q}\Sigma_{gj\in B_Co_Scl} sim_scl(t_i,g_j))$ (7)

- A partir du profil de l'ami U_k :

- $Sim_u_k(B_Co_Them,Q)= (DEP\ (u))*log(\Sigma_{ti\in Q}\Sigma_{tj\in B_Co_Them} sim_them(t_i,t_j))$ (8)

- $Sim_u_k\ (B_Co_Scl, Q)= (DEP\ (u))*log(\Sigma_{ti\in Q}\Sigma_{gj\in B_Co_Scl} sim_scl(t_i,g_j))$ (9)

Dans les formules 6 et 7, en addition de la similarité liant les termes de la requête avec les termes et tags de profil de l'utilisateur, l'indépendance sociale de l'utilisateur est prise en compte. Et dans les deux autres formules 8 et 9, en plus de la similarité des termes de la requête avec les termes et tags des profils des amis proches, la dépendance sociale de l'utilisateur est prise en compte.

Pour alléger l'opération de sélection de la base la plus adéquate à la requête, juste les profils des 10 amis les plus proches (selon la formule 5) qui sont concernés par l'application du formules 8 et 9.

La base choisie comme contexte courant de la recherche est celle la plus similaire à la requête.

3.2. Reformulation de la requête

Le but des techniques de reformulation de requête par réinjection de pertinence [Rocchio], [Efthimiadis 1996],[Tamine 2000]est de générer une nouvelle requête mieux adaptéeau besoin en information de l'utilisateur. La reformulation de requête consisteà ajouter de nouveaux termes à une requête initiale ou alors repondérer lespoids de ses termes dans le but de cibler la recherche vers les documents pertinents. La génération de la nouvelle requête s'effectue soit par reformulationautomatique ou par reformulation interactive.

Respectant ce principe, nous allons enrichir la requête d'utilisateur Q, pour produire une nouvelle requête Q', par l'ajout de termes ou de tags selon le contexte actuel de la recherche exprimé par la base sélectionnée du profil où auprès d'un ami.

Donc, selon la base sélectionnée, la nouvelle requête reformulée Q' est obtenue sur la base de la requête initiale Q comme suit :

$$Q' = Q + P \qquad (10)$$

Avec :

- En cas de B_Co_Them: P est le terme qui maximise la similarité (3) avec les termes de Q.

- En cas de B_Co_Scl: P est le tag qui maximise la similarité (4) avec les termes de Q.

3.3. Mise à jour de profil

La mise à jour d'un profil doit assurer d'un côté, son enrichissement et son évolution dans le temps par l'ajout des nouveaux documents, termes et tags pertinents issus des nouvelles sessions de recherche, et la mise à jour des bases de cooccurrence thématique et sociale. Et d'un autre côté, elle doit veiller sur l'optimisation de l'espace occupé et le temps de réponse pris par les différents calculs, et ce par la suppression des documents, des termes et des tags qui ne figurent pas dans de nouvelles recherche pendant une année. La liste des amis proches est aussi mise à jour périodiquement.

4. Evaluation

4.1. Difficulté d'évaluation dans le cadre de la recherche personnalisée

Selon (Toms et al., 2013; SaoudetKechid, 2016), la pertinence à améliorer dans un cadre de personnalisation de la recherche d'information est trop subjective, de fait qu'elle dépend en particulier de l'utilisateur qui est le seule qui peut juger cette pertinence selon son point de vue. Cette réalité fait de l'évaluation d'une approche de recherche d'information personnalisée une tache très problématique.

Un cadre d'évaluation des approches de RIP doit mettre en place les deux composantes principales suivantes (Bouhini et al. 2016) : les profils des utilisateurs contenant les données caractérisant ces utilisateurs et leurs intérêts, et la collection de testes contenant un ensemble de documents et un ensemble de requêtes avec les pré-jugements de pertinence centré-utilisateur.

Vue l'absence des cadres consensuels dédiés à l'évaluation des approche de la personnalisation de la recherche, les auteurs ont eu recours à concevoir leurs propres solutions (Bouhini et al 2016; SaoudetKechid, 2016; Duong et al. 2013; Daoud et al. 2011 ; …)

4.2. Stratégie de l'évaluation

Vu l'absence d'un cadre d'évaluation dédié au contexte de personnalisation comme c'est déjà indiqué, nous avons construit notre cadre d'évaluation comme suit :

- *Création des profils d'utilisateur* : nous avons construit 20 profils pour 20 utilisateurs appartenant à des domaines professionnels différents pour avoir des centres d'intérêts différents. Nous avons construit chaque profil avec la collaboration de l'utilisateur concerné. Chaque utilisateur fournit un ensemble de documents pertinents de ses dernières recherches, avec les tags qu'il a utilisé pour annoter ces documents. Les documents sont traités pour extraire les termes et les tags pertinents selon les formules proposées, et les liens de cooccurrence sont calculée et sauvegardés dans les deux bases B_Co_Them et B_Co_Scl. Les liens d'amitié sont aussi calculés selon la formule 5 pour établir la liste des amis.

  Les données collectées concernant les utilisateurs sont organisée selon ce qui indique le schéma suivant :

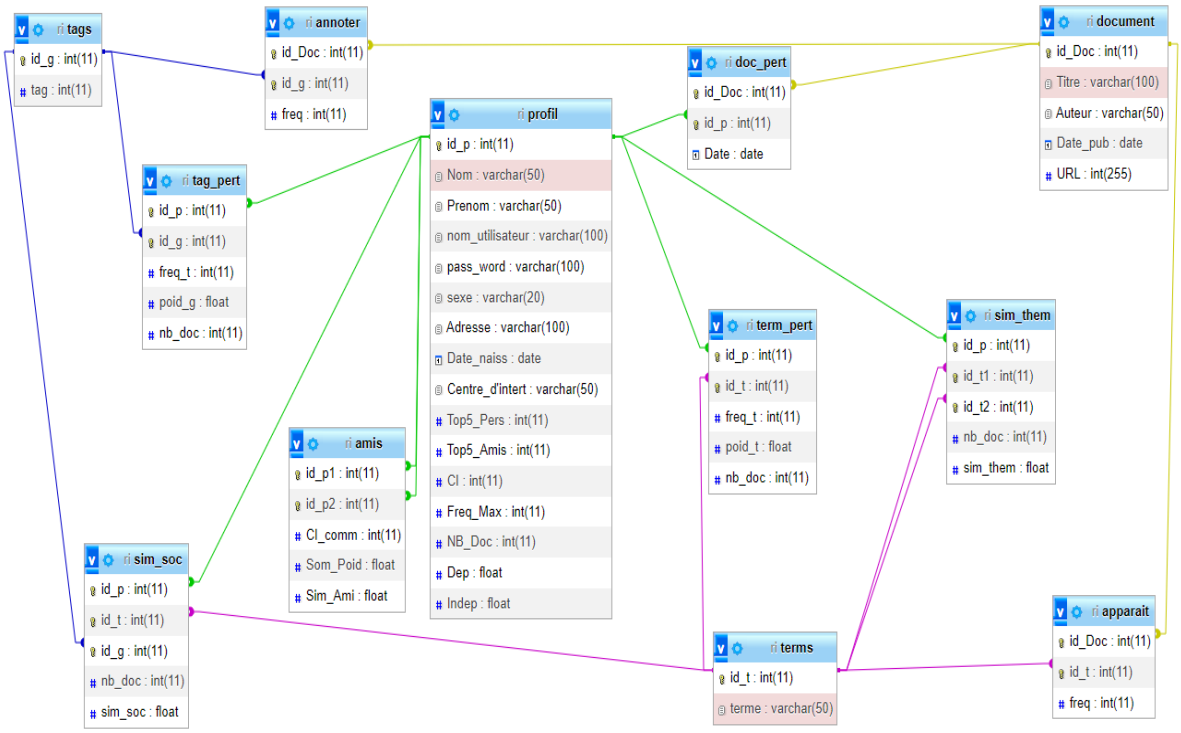


Figure3.6.Représentation des table de la base de donné

- *Besoins en information et requêtes* : pour chaque utilisateur on a défini 5 besoins en information et 5 requêtes qui les expriment. Pour chaque utilisateur, au moins un besoin est définit de façon très spécifique selon son profil et un autre besoin complètement nouveau. Les requêtes sont formulées de façon générale et certaines entre elles sont communes entre les utilisateurs (en somme 100 requêtes différentes).
- *Jugement de pertinence*: les jugements de pertinence des documents sont faits manuellement par les utilisateurs. Chaque document est évalué par (0 : non pertinentou 1 : pertinent).
- *Evaluation* : l'évaluation consiste à calculer la précision (le nombre de document pertinents) dans les 5, les 10 puis les 15 premiers documents retournés. Ce qui représente l'équivalent de 3 premières pages pouvant être visitées à la moyenne par un utilisateur.

5. Conclusion

Dans ce chapitre nous avons présenté notre approche de personnalisation de recherche. Nous avons construit un profil d'utilisateur qui représente l'utilisateur par la combinaison de l'information thématique et sociale. L'information sociale est aussi combinée entre annotations et liens d'amitié.

Une pseudo-sémantique basée sur les liens de cooccurrence des composantes de profil est introduite. Comme nous avons proposé des scores pour mesurer la dépendance et l'indépendance de l'utilisateur concernant sa perception de pertinence.

Conclusion générale

Conclusion Générale

Dans ce travail, nous avons proposé une nouvelle approche pour la personnalisation de la recherche d'information, basée sur la définition d'un profile social qui combine les indicateurs d'intérêts de l'utilisateur issus des contextes thématique et social de ses recherches précédentes. L'aspect social lui-même est combiné par un aspect social interne représenté par les annotations sociales (tags) et l'aspect social externe représenté par les liens d'amitié avec les autres utilisateurs de système de recherche.

Nous avons introduit, pour l'exploitation des termes et tags pertinents du profil utilisateur, une pseudo-sémantique qui s'intéresse aux liens de cooccurrence entre les termes eux-mêmes et entre les termes et les tags. Ainsi que nous avons introduit pour l'exploitation des liens d'amitiés, un score qui mesure la tendance de l'utilisateur d'être influencé par son environnement social ou d'être indépendant en matière de perception de pertinence.

Cette approche est proposée dans l'objectif d'améliorer la pertinence des résultats de recherche retournés, par la reformulation de la requête initiale pour mieux exprimer le réel besoin en information de l'utilisateur.

Ce travail ouvre voie vers d'autres pistes de recherche telles que :

- La réflexion à l'implémentation d'un cadre d'évaluation en ligne dédié à la personnalisation de la recherche d'information, avec des comptes utilisateurs et leurs profils, des collections de tests et des pré-jugements des utilisateurs pour des requêtes prédéfinies, …etc.
- Etudier la stabilité et l'évolution des intérêts de l'utilisateur, pour estimer à quel point son profil peut être utile à inférer ses futurs besoins informationnels.

Bibliographie

Bibliographie

[Abdou et al.2006] : AbdouroihamaneAnli, Mourad Abed PerSyst : Un Système de Personnalisation de l'information transport multimodale. Laboratoire d'Automatique, de Mécanique et d'Informatique industrielles et Humaines UVHC, Le Mont Houy, F-59300 Valenciennes Cedex 9, France, 2006.

[Adamson et Boreham 1974] : Adamson, G. W., &Boreham, J. (1974). The use of an association measure based on character structure to identify semantically related pairs of words and document titles. Information storage and retrieval, 10(7), 253-260.

[Aimé et al. 2010] : Aimé, X, F Fürst, P Kuntz and F Trichet (2010), « Enrichissement sémantique de requêtes au moyen d'ontologies de domaine personnalisées. In Actes de l'atelier 10ieme Journées francophones d'Extraction et de Gestion de Connaissances (EGC'2010), pp. 19–26. Hammamet, Tunisie.

[Akermi et al. 2015] : Akermi, I., Boughanem, M., Faiz, R.: Une approche de recommandation proactive dans un environnement mobile. In: INFORSID 2015, France, pp. 301–316 (2015).

[Amato 1999]: G. Amato, U. Straccia, User Profile Modeling and Applications to Digital Libraries, Proc. 3rd European Conf. Research and Advanced Technology for Digital Libraries, ECDL, 1999.

[Amato et Straccia 1999]: Amato, G. et Straccia, U. (1999). "User profile modeling and applications to digital libraries". In Research and Advanced Technology for Digital Libraries 1999, pp. 184-197.

[Arikan et al 2009]: Arikan, I., Bedathur, S.J., Berberich, K. (2009), "TimeWill Tell: Leveraging Temporal Expressions in IR". In Proceedings of the 2ND International Conference on Web Search and Web Data Mining, WSDM 2009, Barcelona, Spain, ACM (2009).

[Badache et al. 2019]: Badache, I, A Abu-Thaher, M Hamdan and L Abu-Jaish (2019), "Recherched'informationsociale en langue arabe: Cas de facebook", In 16th CORIA'19, Lyon-France.

[Baeza-Yates,Ribeiro-Neto1999] : R. Baeza-Yates and R. A. Ribeiro-Neto. Modern Information Retrieval. New York : ACM Press ; Harlow England : Addison-Wesley, cop., 1999.

[Basil et al. 2014] : PierpaoloBasile, Annalina Caputo, Giovanni Semeraro, and Lucia Siciliani, (2014) "Extending an Information Retrieval System through Time Event Extraction", in 8th International Workshop on Information Filtering and Retrieval, DART 2014, Co-located with XIII AI*IA Symposium on Artificial Intelligence, AI*IA 2014 36-47 CEUR Workshop Proceedings.

[Bendella and Quafafou 19] : Meryem Bendella and Mohamed Quafafou (2019) « Expansion de requêtes à base de motifs et deplongements de mots pour améliorer larecherche de microblogs », 16th CORIA'19, Lyon-France, doi:10.24348/coria.2019.CORIA_2019_paper_16.

[Benjabeur et al. 2010] : Benjabeur, L, L Tamin and M Boughanem (2010), "A social model for Literature Access: Towards a weighted social network of authors", In

International Conference on Adaptivity, Personalization and Fusion of Heterogeneous Information (RIAO 2010), pp. 32–39. France.

[Bila et al. 2008] : Bila, N, J Cao, R Dino, T Ho, R Hull, B Kumar and P Santos (2008), "Mobile user profile acquisition through network observables and explicit user queries", In 9th International Conference on Mobile Data Management, pp. 98–107, Beijing, China.

[Billsus et Pazzani 1999] : Billsus, D., &Pazzani, M. J. (1999), "A hybrid user model for news story classification", In: Kay J. (eds) UM99 User Modeling. CISM International Centre for Mechanical Sciences (Courses and Lectures), vol 407.Springer, Vienna.https://doi.org/10.1007/978-3-7091-2490-1_10.

[Bischoff et al. 2008] : Kerstin BISCHOFF et al. (2008), « Can All Tags Be Used for Search ? » In Proceedings of the 17th ACM Conference on Information and Knowledge Management. CIKM '08. California, USA (2008), p. 193-202. DOI : 10.1145/1458082.1458112.

[BK 05]: M. Bouzeghoub and D. Kostadinov. Personnalisation de l'information : Aperçu de l'état de l'art et définition d'un modèle flexible de définition de profils. In Actes de la seconde édition de la Conférence en Recherche d'Information et Applications(CORIA), pages 201.218, Grenoble, France, 2005.

[Blanco et al. 2008] : Blanco-Fernández, Y., Pazos-Arias, J. J., Gil-Solla, A., Ramos-Cabrer, M., &López-Nores, M. (2008), "Semantic reasoning: A path to new possibilities of personalization", ESWC 2008, LNCS 5021, 2008. pp. 720-735.

[Borlund 2003]: Borlund, P. (2003), "The concept of relevance in IR". Journal of the American Society for Information Science and Technology, 54(10), pp. 913–925.

[Bernardo et al. 2013] : Bernardo Pereira Nunes, Stefan Dietze, Marco A. Casanova ,Ricardo Kawase, BesnikFetahu, and Wolfgang Nejd (2013) "Combining a co-occurrence-based and a semantic measure for entity linking », 10th (ESWC'13) Extended Semantic Web ConferenceAt: Montpellier, FranceVolume: Lecture Notes in Computer Science, Vol. 7882. DOI: 10.1007/978-3-642-38288-8_37.

[Boubkeur 2008] :BOUBEKEUR-AMIROUCHE Fatiha, « Contribution à la définition de modèles de recherche d'information flexibles basés sur les CP-Nets », thèse de doctorat, univ Toulouse 3, France. Juillet 2008.

[Bouhini 2014] : BouhiniChahrazad, « Impact des réseaux sociaux sur le processus de recherche d'information», thèse de doctorat,univ jean Monnet de saint-Etienne, univlyon, France, Octobre 2014.

[Bouhini et al 2016]: Bouhini C, M Gery and C Largeron (2016). "Personalized information retrieval models integrating the user's profile", IEEE 10th International Conference on Research Challenges in Information Science (RCIS 2016), pp. 1–9, Grenoble, France.

[Bouidguaguen 2011] : BouidguaguenOurdia, « Accès contextuel à l'information dans un environnement mobile : approche basée sur l'utilisation d'un profil situationnel de l'utilisateur et d'un profil de localisation des requêtes. », Thèse de doctorat, Université Sabatier, Toulouse, France, decembre 2011.

[Bouramoul et al.2011] :A.Bouramoul, M-K. Kholladi et B-L. Doan : " Evaluation of Information Retrieval Systems Towards a New Context-Based Approach". In

4th.International Conference on Information Systems and Economic Intelligence, IEEE. Marrakech, Morocco, 2011.

[Challam et al 2007] : Challam V., Gauch S., Chandramouli A., « Contextual Search Using Ontology-Based User Profiles », Proceedings of RIAO 2007, Pittsburgh USA, 2007.

[Daoud 2009] : 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 et al. 2008]: Daoud, M., Tamine-Lechani, L., &Boughanem, M. (2008), "Using a graph-basedontological user profile for personalizingsearch". In Proceedings of the 17th ACM conference on Information and knowledge management (pp. 1495-1496).ACM, 2008.

[Daoud et al. 2011]: Daoud, M, L Tamine and M Boughanem (2011), "A personalized search using a semantic distance measure in a graph-based ranking model". Journal of Information Science, 37(6), 614–636, doi: 10.1177/0165551511420220.

[Duong et al. 2013]: Duong, TH, MN Uddin and CD Nguyen (2013), "Personalized semantic search using ODP: A study case in academic domain". In ICCSA 2013, Part V, LNCS 7975, pp. 607–619.

[Efthimiadis 1996]: E. N. Efthimiadis. Query expansion. Annual Review of Information Systems and Technology, 31 :121–187, 1996.

[Fouad et al. 2012]: Fouad, KM, AR Khalifa, NM Nagdy and HM Harb (2012), "Web-based semantic and personalized information retrieval", IJCSI International Journal of Computer Science Issues, 9(3), 266–276. 2012.

[Fox 1992]: Fox, C. (1992, June). Lexical analysis and stoplists.In Information retrieval (pp. 102-130).Prentice-Hall, Inc.

[Haddad 15] : Haddad Lakri, « Un modèle de langue pour une recherche d'information personnalisée », Theèse de Master, Université Mouloud Mammeri, Tizi-Ouzou,Algerie,2015.

[Hattori et al. 2007]: Hattori S., Tezuka T., Tanaka K. (2007), «Context-aware query refinement for mobile web search», Proc. of the 2007 International Symposium on Applications and the Internet Workshops.

[Heymann et al. 2008]: Paul Heymann, Daniel Ramageet Hector Garcia-Molina (2008), « Social Tag Prediction ». In Proceedings of the 31st ACM SIGIR'08, Singapore, DOI: 10. 1145 / 1390334.1390425.

[Kanhabua et Nørvag 2012]: Kanhabua, N., Nørvag, K. (2012), "Learning to Rank Search Results for Time sensitive Queries", In Proceedings of the 21st ACM International Conference on Information and Knowledge Management CIKM '12, pp. 2463– 2466, 2012.

[Kostadinov 2003]: D. Kostadinov. Personnalisation de l'information et gestion des profils utilisateurs, Rapport de DEA, Université de Versailles, France, 2003.

[Kostadinov et al. 2008] : Kostadinov, D., Bouzeghoub, M., & Lopes, S. (2008). "Accès personnalisé à des sources de données multiples. Evaluation de deux approches de reformulation de requêtes ». Ingénierie des Systèmes d'Information, 13(4), 59-82.

[Kumar et al. 2014] : Kumar H., Lee S., G.Kim H., (2014). Exploiting social bookmarking services to build clustered user interest profile for personalized search, Inf. Sci. 281399–417.

[Krir 2012]: HadjouniKrir M. (2012), « Un système de recherche d'information personnalisée basé sur la modélisation multidimensionnelle de l'utilisateur ». Thèse de doctoratUniversité Paris Sud – paris XI, France (2012).

[liu et al. 2004] : Liu F., Yu C., Meng W., « Personalized Web Search For Improving Retrieval Effectiveness », IEEE Transactions on Knowledge and Data Engineering, vol. 16, n˚ 1, p. 28-40, 2004.

[Ma et al. 2007]: Ma Z., Pant G., Sheng, « Interest-based personalized search », ACM Transactions on Information Systems, 2007.

[Marinilli et al 1999]: Marinilli, M., Micarelli, A., &Sciarrone, F. (1999), "A hybrid case-based architecture for information filtering on the web", In ICCBR Workshops (pp. 23-32), University of Kaiserslautern, Computer Science.

[Maron, Kuhns1960]: M. E. Maron and J. L. Kuhns. On relevance, probabilistic indexing and information retrieval. J. ACM, 7(3) :216–244, 1960.

[Mohamed etAbdelmoty 2017]: Mohamed, S and AI Abdelmoty (2017), "Spatio-semantic user profiles in location-based social networks". IN International Journal of Data Science and Analytics, 4, 127–142, 2017. DOI: 10.1007/ s41060-017-0059-9.

[Monteserin et Armentano 2019]: Ariel Monteserin Marcelo and G. Armentano (2019) « Influence me! Predicting links to influential users », Information Retrieval Journal (2019) 22, pp 32–54, https://doi.org/10.1007/s10791-018-9335-0.

[Moukas 1997]: Moukas A., « Information discovery and filtering using a multi-agent evolving ecosystem », Applied Artificial Intelligence, pp. 437±457, 1997.

[Moulahi 2015] : BilelMoulahi (2015), « Définition et évaluation de modèles d'agrégation pour l'estimation de la pertinence multidimensionnelle en recherche d'information », thèse de doctorat, Université Toulouse III Paul Sabatier, 2015.

[Mousset et al. 2019] : P Mousset, Y Pitarch, L Tamine (2019). « Régularisation Spatiale de Représentations Distribuées de Mots", In proceedings of 16th CORIA'19 conference, Lyon, France. Doi:10.24348/coria.2019.CORIA_2019_paper_23.

[Noll et Meinel 2007]: M. G. Noll et Christoph Meinel, (2007) « Web SearchPersonalization via Social Bookmarking and Tagging ». In Proceedings of the 6th International conference of the Semantic Web and 2nd Asian Conference on Asian Semantic Web Conference.2007, p. 367-380.

[Ould Amer 2020] : Nawal OULD AMER, (2020), «Recherche Sociale et Personnalisée d'Information », thèse de doctorat, université de Grenoble Alpes, France.

[Ould Amer et al. 2015] : Nawal Ould Amer, Philippe Mulhem and Mathias Géry (2015), "Recherche de conversations dans les réseaux sociaux : modélisation et expérimentations sur Twitter », 12th CORIA'15 conférence, Paris-France. doi:10.24348/coria.2015.6.

[Panayiotouet Samaras 2006]: Panayiotou C., Samaras G., (2006), «Mobile User Personalization with Dynamic Profiles: Time and Activity», On the Move to Meaningful Internet Systems 2006: OTM 2006 Workshops, LNCS 4278, pp.1295-1304, 2006.

[Peter 1957]: Hans Peter Luhn. A statistical approach to mechanized encoding and searching of literary information.IBM Journal of research and development, 1(4):309–317, 1957.

[Raza et al. 2018]: Muhammad AhsanRaza ,RahmahMokhtar and Noraziah Ahmad (2018) "A survey of statistical approaches for query expansion". IN Journal of Knowledge and Information Systems, published by Springer Nature 2018. https://doi.org/10.1007/s10115-018-1269-8

[Robertson,Sparck Jones 1976] : S. Robertson and K. Sparck Jones. Relevance weighting for search terms. Journal of The American Society for Information Science, 27(3) :129–146, 1976.

[Rocchio]: J. J. Rocchio. Relevance feedback in information retrieval.In Salton.

[Ruthven 2003]: Ruthven, I. (2003). Re-examining the potential effectiveness of interactive query expansion. In Proceedings of the 26th annual international 133 Chapitre 7 ACM SIGIR conference on Research and development in informaion retrieval, SIGIR '03, pages 213–220. ACM.

[Salton 1971] :G. Salton. A comparison between manual and automatic indexing methods. Journal of American Documentation, 20(1) :61–71, 1971.

[Salton 1971]: Gerard Salton. The smart retrieval system—experiments in automatic document processing. 1971.

[Salton et Buckley 1988]: Salton, G. and Buckley, C., (1988) "Term Weighting Approaches in Automatic Text Retrival", Information Processing and Management. 513-523.

[Salton,McGill,1983] : G. Salton and M. McGill. Introduction to Modern Information Retrieval.McGraw-Hill, New York, 1983.

[Saoud 2017] : Saoud Z. « Exploitation des réseaux sociaux pour la recherche d'information », thèse de doctorat, USTHB, Alger, Algérie. Juillet 2017.

[SaoudetKechid 2016]: Saoud, Z and S Kechid (2016), "Integrating social profile to improve the source selection and the result merging process in distributed information retrieval". Information Sciences, 336, 115–128, http://dx.doi.org/10.1016/j.ins.2015.12.012.

[Saracevic 2007]: Saracevic, T. (2007), "Relevance: A review of the literature and a framework for thinking on the notion in information science. part iii : Behavior and effects of relevance", Journal of the American Society for Information Science, 58(13):2126–2144.

[Schenkel et al. 2008]: Schenkel, R., Crecelius, T., Kacimi, M., Michel, S., Neumann, T., Parreira, J. X., and Weikum, G. (2008). "Efficient top-k querying over social-tagging networks". In 31st Conference on Research and Development in Information Retrieval, SIGIR'08, pages 523–530.

[Shafik et al 2015] :O.Shafiq, R.Alhajj, J.G.Rokne, (2015) "On personalizing web search using social network analysis", Inf.Sci. 314(2015)55–76. http://dx.doi.org/10.1016/j.ins.2015.02.029

[Sieg et al2007] : Sieg A., Mobasher B., Burke R., « Web search personalization with ontological user profiles », CIKM'07 : Proceedings of the sixteenth ACM conference on information and knowledge management, ACM, New York, NY, USA, p. 525-534, 2007.
[Smyth 2011] : Smyth, B. (2011). Web search: Social & collaborative, in: Proceedings of 8th French Information Retrieval Conférence en Recherched'Infomations et Applica-tions, Avignon, France, March16–18, ,CORIA,2011,p.3.
[Tamine 2000] : L. Tamine. Optimisation de requêtes dans un système de recherche d'information : approche basée sur l'exploitation de techniques avancées de l'algorithmique génétique. Thèse de doctorat en informatique, Université Paul Sabatier, decembre 2000.
[Tamine et Daoud 2018] :Tamine, L and M Daoud (2018). « Evaluation in contextual information retrieval: Foundations and recent advances within the challenges of context dynamicity and data privacy". ACM Computing Surveys, 51(4), 1–36, https://doi.org/10.1145/3204940.
[Tamine et al. 2008]: Tamine L., Boughanem M., Zemirli W. N., « Personalized document ranking : Exploiting evidence from multiple user interests for profiling and retrieval », Journal of Digital Information Management, vol. 6, n° 5, p. 354-365, octobre, 2008.
[Taylor et al. 2007]: Taylor, A. R., Cool, C., Belkin, N. J. etAmadio, W. J. (2007), "Relationships between categories of relevance criteria and stage in task completion". Information Processing and Management, 43(4):1071–1084.
[Tchuente 2013] : Tchuente Dieudonné (2013), " modélisation et dérivation des profils utilisateurs à partir des réseaux sociaux : approche à partir de communauté de réseaux k-egocentriques ", thèse de doctorat, Université Paul Sabatier - Toulouse III, France, 2013.
[Toms et al. 2013]: Toms, E, M Agosti, N Fuhr and P Vakkari (2013). Evaluation methodologies in information retrieval.Dagstuhl Reports, 3(10), 92–126.
[Uddin et al. 2012]: Uddin, MN, TH Duong, V Sean and G-S Jo (2012), "Construction of semantic user profile for personalized web search". In ICCCI 2012, Part II, LNAI7654, pp. 99–108. Springer.
[Vallet et al. 2010] : David Vallet, IvánCantador et Joemon M. Jose. « Personalizing Web Search with Folksonomy-Based User and Document Profiles ».In Advances in Information Retrieval, 32nd ECIR 2010, UK, p 420-431.
[VandenbusscheetTeissèdre 2011]: Vandenbussche, P.Y., Teissèdre, C., (2011), "Events Retrieval Using Enhanced SemanticWeb Knowledge". In Workshop DeRIVE2011, in conjunction with 10th International Semantic Web Conference 2011 (ISWC 2011).
[Xie et al. 2012] : Hao-Ran XIE, Qing LI et Yi CAI. (2012), « Community-Aware Resource Profiling for Personalized Search in Folksonomy ». Journal of Computer Science and Technology 27(3): 599–610 (2012), DOI: 10.1007/s11390-012-1247-7.
[Xu et al. 2008] : Shengliang XU et al. (2008) « Exploring Folksonomy for Personalized Search ». In Proceedings of the 31st ACM SIGIR'08, Singapore, URL: https://doi.org/10.1145/1390334. 1390363.

[Zargayouna et al. 2015]: Zargayouna, H, C Roussey and J-P Chevallet (2015). "Recherche d'information sémantique: état des lieux ». Traitement Automatique des Langues TAL, 56(3), p 49–73.

[Zemirli 2008] : Zemirli W.N, « Modèle d'accès personnalisé a l'information basé sur les diagrammes d'influence intégrant un profil multidimensionnel », Thèse de doctorat, Université Paul Sabatier, Toulouse, France, juin 2008.

[ZTB, 05] :W. N. Zemirli, L. Tamine, and M. Boughanem. Accès personnalisé à l'information : vers la définition d'un profil utilisateur multidimensionnel. In International Symposium On Programming Systems, pages 20.28. USTHB, 2005.