Impact des réseaux sociaux sur le processus de recherche d'information

Chahrazed Bouhini

Thèse présentée pour obtenir le grade de docteur

❧

Laboratoire Hubert Curien, UMR CNRS 5516
Faculté des Sciences et Techniques, Université Jean Monnet de Saint-Étienne, Universités de Lyon

Discipline : Informatique

Impact des réseaux sociaux sur le processus de recherche d'information

Par : **Chahrazed Bouhini**

Sous la direction de Christine Largeron et Mathias Géry,

Membres du jury:

Sylvie Calabretto, Professeur, INSA de Lyon, **Rapporteur**

Philippe Mulhem, Chargé de Recherche CNRS - HDR, LIG, **Rapporteur**

Florence Sèdes, Professeur, Université Paul Sabatier de Toulouse, **Examinatrice**

Michel Beigbeder, Maître Assistant, ENSM de Saint-Étienne, **Examinateur**

Christine Largeron, Professeur, Université Jean Monnet, **Directrice de thèse**

Mathias Géry, Maître de conférence, Université Jean Monnet, **Co-encadrant**

Date de soutenance : 21 Octobre 2014

Remerciements

Je remercie tout d'abord mes encadrants de thèse, ma directrice de thèse Christine Largeron et mon co-encadrant Mathias Géry qui ont été toujours disponibles et n'ont jamais hésité à me consacrer leurs temps et leur énergie et m'ont appris des tas de choses au cours de ma thèse dont je suis vraiment et sincèrement reconnaissante, sans leurs efforts et détermination ce travail n'aurait pas pu aboutir.

Je tiens à exprimer ma gratitude et mes remerciements les plus sincères aux membres de jury de m'avoir fait l'honneur d'accepter de faire partie de mon jury de thèse. Un grand merci à mes rapporteurs : Sylvie Calabretto, Phillipe Mulhem et ma présidente de jury Florence Sédes que j'ai eu souvent l'opportunité et l'immense plaisir de les rencontrer et échanger avec eux lors des différentes conférences. Enfin, je remercie également Michel Beigbeder mon professeur à l'école des mines de Saint-Etienne de m'avoir initier à la recherche d'information et pour bien avoir voulu être examinateur de ma thèse.

Je remercie aussi tous mes collègues au laboratoire Hubert Curien et à l'université Jean Monnet, plus particulièrement mes collègues et anciens collègues de bureau : JP, Fabien, David, Tung, Aurélien, Laurent, Emilie, Mattias, Nidhal, Rahaf, Michael, Taygun sans oublier mes collègues de l'université d'Angers pour leurs support et encouragements.

Chaleureusement je remercie ma famille bien aimée, plus particulièrement mes très chers parents pour tout ce qu'ils ont fait et continuent de faire pour moi et qui sont à l'origine de ma réussite sur tous les plans, je n'en serai jamais arrivée sans leur amour et leur soutien durant mes meilleurs et mes pires moments de ma vie. Les seules personnes qui n'ont jamais cessé de croire en moi même dans mes moments de doutes.

A toutes les personnes très chères à mon cœur et tous mes nombreux ami(e)s qui se reconnaîtront, merci d'avoir été toujours là pour moi et d'avoir cru en moi. Enfin, un très grand merci à tous mes enseignants et professeurs des universités et écoles que j'ai fréquentées et à qui je dédier ma citation préférée d'Albert Einstein "*It is the supreme art of the teacher to awaken joy in creative expression and knowledge*", grâce à vous tous j'ai pu atteindre mes objectifs et réaliser une bonne partie de mes rêves.

Vifs remerciements à tous ceux qui ont contribué de près ou de loin à l'élaboration de ce travail. Sur le plan personnel comme sur le plan professionnel, à toute personne m'ayant appris la moindre chose, merci.

Table des matières

Table des figures

Liste des tableaux

Résumé

L'émergence des réseaux sociaux a révolutionné le Web en permettant notamment aux individus de prolonger leur connexion virtuelle en une relation plus réelle et de partager leurs connaissances. Ce nouveau contexte de diffusion de l'information sur le Web peut constituer un moyen efficace pour cerner les besoins en information des utilisateurs du Web, et permettre à la recherche d'information (RI) de mieux répondre à ces besoins en adaptant les modèles d'indexation et d'interrogation.

L'exploitation des réseaux sociaux confronte la RI à plusieurs défis dont les plus importants concernent la représentation de l'information dans un modèle social personnalisé de RI et son évaluation, en l'absence de collections de test et de compétitions dédiées.

Nous proposons dans ce travail de bénéficier de l'exploitation des informations issues des réseaux sociaux pour personnaliser la recherche d'information de l'utilisateur en se rapprochant le plus de ses centres d'intérêt et de ses préférences.

Les principales contributions de notre travail consistent dans un premier temps à établir un profil social de l'utilisateur à partir du contenu informationnel généré au sein du réseau social. Nous présentons par la suite des modèles de recherche sociale personnalisée d'information (RSPI) permettant d'intégrer le profil social de l'utilisateur à différents niveaux du processus de RI. Dans l'objectif de permettre l'évaluation des modèles de RSPI sur une collection de test dédiée, nous proposons une collection de test de RSPI que nous avons construite à partir du réseau d'annotation collaborative "*Delicious*" [1], contenant en plus des données classiques d'une collection de test de RI, des données centrée-utilisateur (requêtes centrées utilisateur et jugements de pertinence par requête centrés utilisateur).

1. Delicious : http ://delicious.com

Abstract

The emergence of social media has revolutionized the web by allowing individuals to extend their virtual connection in a more real relationship and share knowledge. This new context of information dissemination on the Web can be an effective way to identify the information needs of Web users, and allow information retrieval (IR) to better meet these needs by adapting the indexing and querying models. The information retrieval faced several challenges with the use of social networks, the most important concerns the representation of information in a personalized social IR (PSIR) model and its evaluation in the absence of a social test collections with the user-centered data (user-centered queries and user-centered relevance judgments). We propose to benefit from the use of the user generated content (UGC) on the social networks to personalize his social search in order to better fit his interests and preferences. The main contributions of our work consist of, on the one hand, building a social profile from the UGC within the social network. We propose then a personnalized social information retrieval models which integrate the user's social profile at various levels of the IR process. On the other hand, with the objective of evaluating our PSIR models on a dedicated test collection, we propose a PSIR test collection "DelRSI" we built from the collaborative social bookmarking network "Delicious" ; a PSIR test collection containing in addition to the classical IR test collection's data, a user-centered data.

Chapitre 1

Introduction

1 Contexte

Le nombre d'internautes ne cesse de grandir : les dernières statistiques établies en 2012 par *Internet World Stats*[1] montrent que le nombre d'internautes a atteint les 2.500 millions, l'équivalent de 34,3% de la totalité de la population du monde entier comme montré dans la figure 1.1[2].

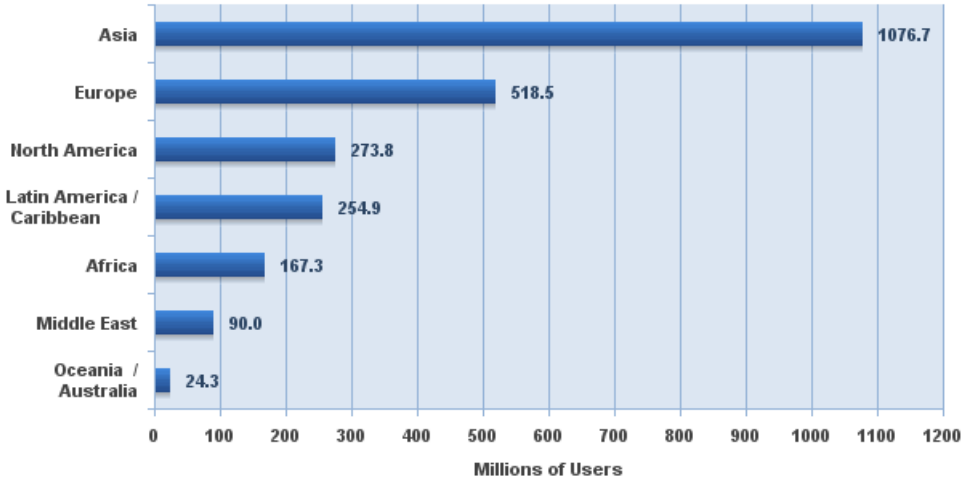

Figure 1.1 – Les utilisateurs d'internet dans le monde, répartis par région géographique

Avec cette explosion du nombre d'internautes et de la quantité d'informations disponibles sur le Web, différents types de systèmes d'accès à l'information ont vu le jour, notamment les systèmes de recommandation, de filtrage ou de recherche d'information.

1. `http://www.internetworldstats.com/` URL consultée la dernière fois le 09/07/2014
2. Miniwats Marketing Group : `http://www.internetworldstats.com/`

Dans l'objectif d'améliorer les résultats retournés par ces systèmes, deux grands procédés viennent enrichir ces systèmes : la collaboration et la personnalisation.

L'aspect collaboratif des systèmes permet aux utilisateurs d'effectuer différentes tâches de façon collaborative comme il permet aux systèmes de tirer profit de cette collaboration entre utilisateurs en terme d'expertise.

Tandis que l'aspect collaboratif se focalise sur l'apport collaboratif des utilisateurs effectuant les mêmes tâches, la personnalisation, elle, se focalise sur les informations disponibles à propos d'un utilisateur pour personnaliser les résultats qui lui sont retournés selon son profil.

1.1 Principaux systèmes d'accès à l'information

1.1.1 Systèmes de recommandation

Les systèmes de recommandation permettent de recommander à l'utilisateur des contenus sur le Web, souvent appelés items ou produits (pages Web, images, films, musique, livres, etc.), susceptibles de l'intéresser [Goh and Foo, 2008]. Comme défini par Konstas et *al.*, le système de recommandation se sert des informations rassemblées à propos des descriptions des items et des profils des utilisateurs dans l'objectif de retourner à l'utilisateur des items de manière personnalisée en se basant sur le comportement antérieur de l'utilisateur en question [Konstas et al., 2009].

1.1.2 Systèmes de filtrage collaboratif

Le filtrage des informations consiste à faire parvenir à l'utilisateur des informations pertinentes à partir d'un flux d'informations. Pour cela, un système de filtrage collaboratif qui se base sur l'utilisation des informations à propos des individus dont le profil est similaire à celui de l'utilisateur à un instant donné, afin de lui recommander des items [Goh and Foo, 2008].

1.1.3 Systèmes de recherche d'information

La recherche d'information (RI) est le domaine dans lequel s'inscrit cette thèse. Elle vise à définir des modèles et des processus dont le but est de retourner, depuis un corpus de documents indexés, ceux dont le contenu correspond le mieux au besoin d'information exprimé par un utilisateur sous la forme d'une requête.

Plutôt vue comme une science de la recherche dans les documents et initialement développée dans le cadre de corpus de documents textuels, la RI a évolué avec l'émergence du Web et plus récemment des réseaux sociaux.

2 Problématique

Avec l'avènement du Web 2.0 et l'apparition d'une part de la notion de communauté de pratique (*community of practice*), et d'autre part des réseaux sociaux, une dimension sociale vient enrichir les contenus des ressources sur le Web.

Initialement introduite par Wenger, l'idée dans la notion de communauté de pratique est qu'une personne peut mieux satisfaire ses besoins d'information si elle est intégrée dans une communauté de praticiens ayant des intérêts et des problèmes similaires [Wenger, 1996].

Au sein des médias sociaux, les utilisateurs interagissent avec le contenu des ressources sur le Web (pages Web, images, vidéos, etc.) et génèrent une grande quantité d'informations.

Avec l'apparition de cette nouvelle dimension, pour satisfaire les utilisateurs, il y a eu besoin d'adapter les systèmes et les processus (systèmes de recommandation, de filtrage ou de recherche d'information, etc.) en tenant compte des informations issues des réseaux sociaux. Nous considérons en effet que les informations sociales peuvent s'avérer d'une grande importance et qu'elles peuvent en particulier permettre d'identifier et de représenter les préférences et les centres d'intérêt de chaque utilisateur.

L'adaptation de la recherche classique d'information à cette nouvelle dimension sociale, a entraîné l'émergence de la Recherche Sociale d'Information (RSI) et de la Recherche Sociale et Personnalisée d'Information (RSPI).

La RSI permet d'intégrer et d'exploiter diverses informations sociales aux modèles et aux processus de la RI.

La RSPI repose sur le postulat qu'au sein d'un réseau social, deux utilisateurs posant une même requête mais avec des centres d'intérêts différents, n'ont pas forcément le même besoin d'information et donc les mêmes attentes en terme de listes de résultats [Harter, 1992], [Mizzaro, 1997].

En s'inspirant de la fameuse citation de *William Sanford Bill Nye*[3] "*Every one*

3. William Sanford ("The Science Guy") : http ://www.billnye.com/

you will ever meet knows something you don't", nous supposons qu'il est important de tenir compte de l'influence sociale et de son impact sur l'information recherchée par l'utilisateur. Nous considérons qu'un système de RSPI doit être en mesure de retourner des listes de documents en fonction des préférences et des centres d'intérêts de chaque utilisateur.

De nombreux travaux de recherche ont porté sur l'exploitation et l'étude d'informations à propos des utilisateurs issues des réseaux sociaux, et se sont principalement focalisés sur l'analyses des activités des utilisateurs et de leurs comportements afin de mieux cerner leur profil [Al-Khalifa and Davis, 2006], [Michlmayr and Cayzer, 2007], [Au-Yeung et al., 2008], [Stoyanovich et al., 2008], [Szomszor et al., 2008], [Yamaguchi et al., 2010], [Vallet et al., 2010], [Cai and Li, 2010], [Liu et al., 2013a], [Liu et al., 2013b], [Lagnier et al., 2013].

Les questions qui se posent généralement en RSPI, sont, premièrement, de savoir ce que représente ce genre d'information pour l'utilisateur et sa recherche d'information, et deuxièmement, comment intégrer cette nouvelle source d'information au sein du processus de RI , et en particulier comment la combiner avec les informations d'un modèle classique de RI (requête et ressources documentaires sous forme d'images, de pages web, de vidéos, etc.).

En effet, les informations sociales peuvent porter sur les préférences de l'utilisateur et les thématiques qui l'intéressent, mais aussi sur les ressources sur le Web. Dans notre travail, nous nous focalisons sur l'exploitation des informations sociales comme étant des informations sur les préférences et les thématiques intéressant l'utilisateur, afin de mieux cerner les besoins d'information des utilisateurs et permettre au système de RSPI de personnaliser la recherche et de retourner une liste de documents pertinents qui se rapprochent le mieux de leurs attentes.

3 Exemple illustratif

Afin d'illustrer la RSPI et le besoin de personnalisation des utilisateurs au sein des réseaux sociaux, nous présentons un exemple simple basé sur deux utilisateurs (Alice et Bob) au sein d'un réseau social :

Alice et Bob souhaitent chacun obtenir des informations à propos de *smartphones* et d'*Android*. Ainsi, Alice et Bob posent la requête suivante :

– Requête : *Smartphone Android*

Alice s'intéresse aux smartphones sortis récemment car elle souhaite acheter un nouveau smartphone avec de bonnes caractéristiques et plus particulièrement un smartphone à base d'Android. Bob est beaucoup plus intéressé par le système Android et le développement des applications Android. Il cherche de la documentation sur le système Android utilisé dans les smartphones.

Le besoin d'information d'Alice est donc différent de celui de Bob. Malgré cela, Alice et Bob ont formulé la même requête textuelle. Nous supposons que nous avons deux documents d_1 et d_2 contenant un nombre d'occurrences différents des termes smartphones et Android, comme présenté dans la table 1.1.

Termes	*Smartphone*	*Android*
Requête	1	1
Document d_1	85	7
Document d_2	7	85

Tableau 1.1 – Nombre d'occurrences des termes dans la requête et dans les documents

Sur la base de cette distribution de termes, un système de RI classique retournera la même liste de documents pertinents d_1 et d_2 dans le même ordre pour Alice et pour Bob, même si leur besoin d'information n'est pas le même.

Or, Alice a employé le terme Smartphone 40 fois et 3 fois le terme Android dans ses annotations sociales alors que Bob les a utilisés respectivement 3 fois et 40 fois (cf. tableau 1.2).

De plus, au sein des voisinages sociaux de Alice et de Bob, les amis d'Alice ont employé le terme Smartphone 55 fois et 12 fois le terme Android et les amis de Bob ont utilisé 15 fois le terme Smartphone pour annoter les documents et 54 fois le terme Android.

	Smartphone	Android
Utilisateur Alice	40	3
Amis de Alice	55	12
Utilisateur Bob	3	40
Amis de Bob	15	54

Tableau 1.2 – Nombre d'occurrences des termes dans les annotations sociales

4 Contribution

Compte tenu des préférences et centres d'intérêts d'Alice et de Bob à travers leurs annotations et celles de leurs voisins sociaux (amis), nous faisons l'hypothèse que les documents parlants plus de Smartphone que d'Android seront susceptibles d'intéresser Alice beaucoup plus que Bob et vice versa, les documents qui parlent beaucoup plus d'Android que de Smartphone intéresseront Bob plus qu'Alice.

La personnalisation de la RSI doit permettre au système de RSPI de retourner à Alice d_1 en premier et d_2 ensuite, tandis que d_2 doit être retourné à Bob en premier.

L'objectif de ce travail de thèse est de modéliser les informations sociales à propos des utilisateurs au sein des réseaux sociaux pour pouvoir les incorporer au niveau d'un modèle de recherche sociale personnalisée d'information (RSPI) afin d'adapter les résultats retournés à chaque utilisateur selon ses préférences et ses centres d'intérêt.

Les deux principales contributions de ce travail sont la proposition de modèles de RSPI et la construction d'une collection de test dédiée à la RSPI.

Concernant les modèles de RSPI, nous nous sommes plus particulièrement intéressé à deux aspects :

1. Modélisation du profil de l'utilisateur : où nous proposons d'exploiter des informations sociales côté utilisateur pour la modélisation de ses centres d'intérêt et la construction de son profil social que l'on appellera "contexte informationnel social" (CIS).
2. Intégration du contexte informationnel social de l'utilisateur au sein du modèle de RSPI.

Pour pallier les insuffisances des collections de test construites et afin de permettre l'évaluation des modèles de RSPI nous proposons une approche méthodologique pour la construction d'une collection de test de RSPI avec des requêtes et des jugements de pertinence centrés utilisateur. La collection est construite à partir des annotations extraites depuis le réseau social d'annotation collaboratif *Delicious*.

5 Plan

Dans la suite, nous allons présenter dans le chapitre 2 les principales parties et composantes de la RI et de ses principaux modèles, ainsi que les informations sociales

issues des réseaux sociaux et les problématiques soulevées dans l'exploitation de ces informations au niveau du modèle de RI, pour une recherche sociale personnalisée d'information.

Dans le chapitre 3, nous présentons des travaux d'état de l'art liés à notre travail, portant sur l'exploitation des informations sociales pour la modélisation des centres d'intérêts de l'utilisateur et les différents niveaux d'intégration de ces informations sociales au sein d'un modèle de RSPI, ainsi que sur la construction des différents éléments d'une collection de test de RSPI.

Dans le chapitre 4, nous détaillons notre modèle de RSPI et notre approche de construction d'un contexte informationnel social de l'utilisateur.

Dans le chapitre 5, nous présentons notre approche pour la construction d'une collection de test de RSPI, incluant les différentes étapes de génération automatique des requêtes et des jugements de pertinence centrée utilisateur à partir des annotations sociales du réseau d'annotation collaboratif *Delicious*. Nous montrons aussi quelques résultats d'évaluation d'un modèle classique de RI (BM25) sur cette collection.

Dans le chapitre 6, nous présentons les expérimentations que nous avons menées en détaillant les résultats des modèles de RSPI proposés dans le chapitre 4. Ces modèles sont évalués sur notre collection de test de RSPI détaillée dans le chapitre 5.

Nous terminons par une conclusion et une présentation de quelques perspectives de ce travail de thèse.

$R_S \ : < U, Rel, T, D, A >$	Modèle du réseau social
$U = \{u_1, u_2, ..., u_x, ..., u_{\|U\|}\}$	Utilisateurs du réseau
$Rel \subset U \times U$	Relations sociales au sein du réseau
$T = \{t_1, t_2, ..., t_j, ..., t_{\|T\|}\}$	Ensemble des termes d'index
$D = \{d_1, d_2, ..., d_i, ..., d_{\|D\|}\}$	Collection de documents sur le Web
$A = \{a_1, a_2, ..., a_z, ..., a_{\|A\|}\}$	Liste des annotations des documents
$a_z = < d, u, T_z >$	Annotation associée au document d par l'utilisateur u avec un sous ensemble de termes T_z de T
$A_u = \{a_1, a_2, ..., a_z, ..., a_{\|A_u\|}\}$	l'ensemble des annotations de l'utilisateur u
$Q = \{q_1, q_2, .., q_l, ...q_{\|Q\|}\}$	Ensemble de requêtes
$q = \{t \in T\}$	Requête composée d'un ou plusieurs termes
$Q_{CU} = \{(q, u) \subset Q \times U\}$	Ensemble de couples de requêtes par utilisateur
$Qrels_q = \{qrels_{q_1}, .., qrels_{q_l}, .., qrels_{\|Q\|}\}$	Ensemble de jugements de pertinence globale
$qrels_q \subseteq D$	Ensemble de documents pertinents pour la requête q
$qrels_{q,u} \subseteq D$	Ensemble de documents pertinents pour la requête q et l'utilisateur u
$V(u) = \{u' \ / \ (u, u') \in Rel\}$	Voisinage social de l'utilisateur u
$PS(u)$	Profil social de l'utilisateur u contenant les termes des annotations de u
$PV(u)$	Profil du voisinage social de u composé de l'ensemble des profils des utilisateurs dans le voisinage de u
$w_{d,t} \in \mathbb{R}$	Poids d'un terme $t \in T$ dans le document $d \in D$
$tf_{d,t} \in \mathbb{N}$	Fréquence d'un terme $t \in T$ dans le document d
$w_{u,t} \in \mathbb{R}$	Poids d'un terme $t \in T$ dans le profil $PS(u)$ de l'utilisateur u
$tf_{u,t} \in \mathbb{N}$	Fréquence d'un terme $t \in T$ dans le profil $PS(u)$ de l'utilisateur u
$tf_{V_u,t} \in \mathbb{N}$	Fréquence d'un terme $t \in T$ dans le profil $PV(u)$ de l'utilisateur u
$ws_{d,u,t} \in \mathbb{R}$	Poids d'un terme $t \in T$ dans le document $d \in D$ pour l'utilisateur u

Tableau 1.3 – Tableau des notations

Chapitre 2

Recherche d'information et réseaux sociaux

1 Recherche d'information

L'objectif de la recherche d'information (RI) est de permettre à l'utilisateur un accès facile à l'information pertinente répondant à son besoin d'information. Un système de recherche d'information (SRI) doit d'abord déterminer la nature exacte des besoins d'information de l'utilisateur, puis sélectionner un sous-ensemble de documents qui peuvent satisfaire ses besoins d'information et enfin les classer par ordre décroissant de pertinence.

1.1 Modèle de RI

Les principales étapes d'un processus de RI sont présentées dans la figure 2.1. Un utilisateur formule son besoin d'information par une requête, celle-ci est alors interprétée par le système pour être représentée à l'aide d'un modèle de requête. En parallèle, les documents du corpus sont indexés pour être représentés à l'aide d'un modèle de document.

À l'étape d'indexation du corpus de documents et de la requête, les termes de l'index sont pondérés en fonction de leur distribution au sein du corpus et au niveau de la requête. À l'étape d'interrogation, le système de RI calcule un score de pertinence système pour chaque document, par le biais de la fonction de correspondance entre les termes de la requête et ceux du corpus de documents.

Le système retourne ensuite une liste de documents considérés comme pertinents par rapport à la requête utilisateur en fonction de leur score.

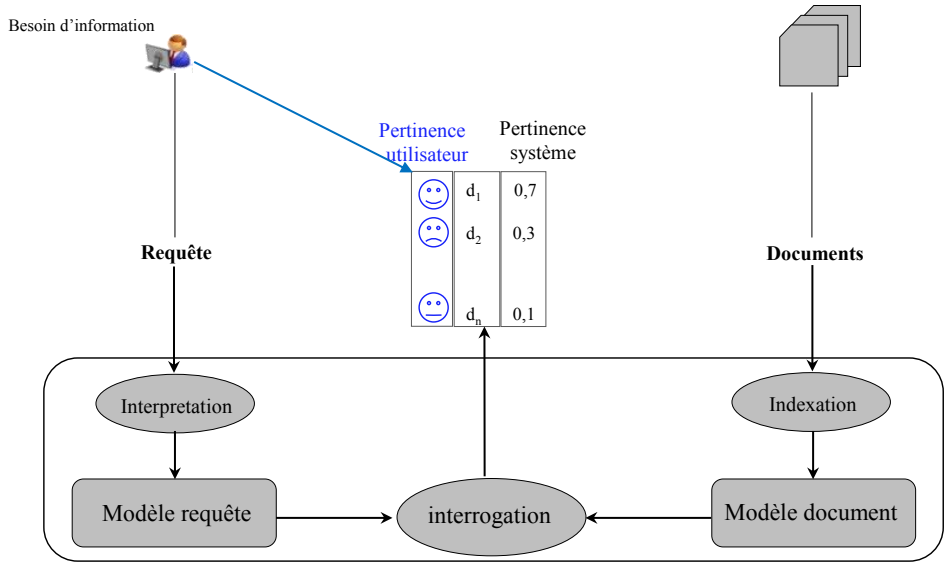

Figure 2.1 – Modèle de recherche d'information

1.1.1 Indexation / interprétation

L'indexation des données textuelles de la requête de l'utilisateur et du corpus de documents constitue le champ de nombreuses études [Sparck Jones, 1974], [Salton, 1986], [Lewis and Croft, 1990]. Elle représente l'une des principales étapes de RI. En effet, indexer un document permet de mettre en avant les termes représentatifs qui le composent et de référencer les documents qui contiennent ces termes dans l'objectif de générer la liste des termes d'indexation.

Ces termes d'indexation seront ajoutés à l'index de la collection avec la liste des références de chaque document les contenant. Un autre objectif de l'indexation peut être d'éliminer les mots vides ayant uniquement un rôle syntaxique et étant donc sans intérêt informationnel.

En plus de l'élimination des mots vides, d'autres techniques d'analyse lexicale sont utilisées lors de l'indexation telles que la **lemmatisation** permettant de regrouper les différentes formes des mots d'une même famille en les réduisant à des mots appelés lemme (forme canonique) et la **racinisation** appelée souvent désuffixation permettant de transformer des flexions en leur racine (ou radical) [Manning et al., 2008].

1.1.2 Pondération

La pondération d'un terme d'indexation est l'association d'une valeur numérique à ce terme de manière à indiquer sa représentativité, estimée à travers la fréquence d'apparition du terme au sein d'un document et son pouvoir de discrimination dans le corpus, obtenu par le biais de la fréquence d'apparition globale du terme au sein de tout le corpus [Manning et al., 2008].

Salton et Buckley décrivent et comparent différentes fonctions de pondération qui permettent de distinguer diffèrents modèles de représentation comme par exemple : le modèle binaire, le modèle vectoriel, etc. [Salton and Buckley, 1988].

1.1.3 Fonction de correspondance et calcul de score

La fonction de correspondance consiste à établir une comparaison entre le document et la requête, ce qui revient généralement à calculer un score mesurant la similarité entre le document et la requête. Ce score de similarité entre le document et la requête est donné par une fonction de correspondance nommée *Retrieval Status Value* [Manning et al., 2008].

Cette fonction permet de comparer le contenu de la requête à chacun des documents indexés [Manning et al., 2008].

Dans la plupart des modèles classiques de RI, le SRI utilise cette fonction pour retourner à l'utilisateur une liste de documents classée par valeur décroissante de pertinence système ($RSV(d, q)$).

1.2 Principaux modèles de recherche d'information

Dans cette partie, nous présentons les principaux modèles de RI, à commencer par les modèles booléens et booléens étendus, puis les modèles vectoriels et les modèles probabilistes.

1.2.1 Modèles booléens

Un des premiers modèles de RI, le modèle booléen, est basé sur la théorie des ensembles [Salton, 1969]. La requête peut être représentée par un ensemble d'un ou plusieurs termes. La particularité du modèle booléen est que les termes de la requête

sont reliés entre eux par des opérateurs de la logique booléenne définis par Georges Boole [1] : OR, AND, NOT.

Un document est également représenté par un ensemble de termes. Par exemple, $d = \{t_1, t_2, ..., t_n\}$. Dans les modèles booléens, l'indexation des documents permet de représenter les termes avec des poids de 0 ou de 1 (1 si le terme apparaît dans le document et 0 sinon).

La pertinence système d'un document par rapport à une requête est ensuite mesurée par un score binaire : $RSV(d,q) = \begin{cases} 1 : si\ \ le\ \ document\ \ est\ \ pertinent \\ 0 : sinon. \end{cases}$

Cette correspondance ($RSV(d,q)$) entre une requête q réduite à un terme t et un document d est déterminée de la façon suivante :

$$\text{RSV(d, t)} = \begin{cases} 1 : si\ \text{t} \in d \\ 0 : sinon. \end{cases}$$

Dans le cas de requêtes composées de plusieurs termes, le score de pertinence système est calculé de la manière suivante :

$$\text{RSV(d,}\quad \text{q}_1\ \ AND\ \ q_2) = \begin{cases} 1 : si\ \text{RSV(d, q}_1) = 1\ \text{ET}\ RSV(d,q_2) = 1 \\ 0 : sinon. \end{cases}$$

$$\text{RSV(d,}\quad \text{q}_1\ \ OR\ \ q_2) = \begin{cases} 1 : si\ \text{RSV(d, q}_1) = 1\ \text{OU}\ RSV(d,q_2) = 1 \\ 0 : sinon. \end{cases}$$

$$\text{RSV(d,}\quad \text{NOT}\ \ \text{q}_1) = \begin{cases} 1 : si\ \text{RSV(d, q}_1) = 0 \\ 0 : sinon. \end{cases}$$

Un des prototypes classiques de moteur de recherche basé sur le modèle booléen est le système *Westlaw* [2] qui est appliqué au cas de la recherche d'information dans des textes de lois.

Cependant, il est difficile pour un utilisateur non expert dans l'utilisation du modèle booléen de formuler des requêtes cohérentes par le biais des expressions booléennes [Belkin and Croft, 1992].

De plus, dans l'approche booléenne, les documents qui satisfont la requête ne sont pas classés et sont tous identiquement similaires à la requête.

Pour remédier aux inconvénients du modèle booléen, le modèle booléen étendu a été proposé en permettant d'indexer les documents d'une manière plus souple, en associant un poids à chaque terme d'indexation.

De cette manière, le système sera en mesure de retourner les documents sur

1. http ://plato.stanford.edu/entries/boole/
2. www.westlaw.com

la base d'un appariement approché (les documents qui se rapprochent le plus de la requête sont retournés dans l'ordre de leur similitude à la requête) au lieu d'un appariement exact (existence ou absence d'un terme dans le modèle booléen de base) [Salton et al., 1983].

1.2.2 Modèles vectoriels

Dans le modèle vectoriel (Vector Space Model "*VSM*"), un document est représenté par un vecteur à $|T|$ dimensions et $|T|$ est l'ensemble des termes de l'index [Salton et al., 1983], [Salton, 1986].

La requête est exprimée sous forme d'un ensemble de mots clés. Les documents et la requête sont représentés comme des vecteurs dans l'espace vectoriel $\mathbb{R}^{|T|}$ (cf. figure 2.2).

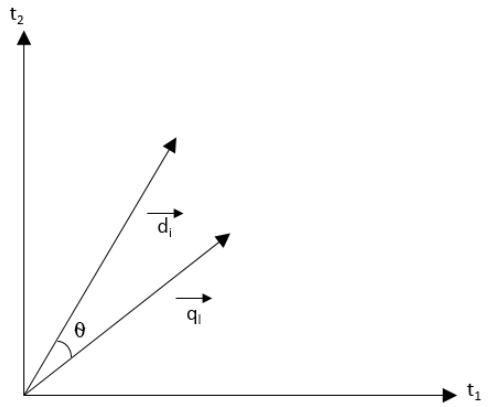

Figure 2.2 – Modèle vectoriel "VSM"

La pondération des termes dans le modèle vectoriel permet d'affecter un poids $w_{d,t} \in [0, 1]$ à chaque terme t.

La fonction de correspondance évalue ensuite la similarité de d par rapport à la requête q pour calculer le score d'appariement entre la requête et le document.

Le score de pertinence système de chaque document est proportionnel à la distance entre les représentations vectorielles du document et de la requête. Plus la distance entre les deux vecteurs est petite plus le score de pertinence du document pour la requête est grand (cf. figure 2.2).

Dans le cas d'un appariement avec un modèle vectoriel à base d'un cosinus, le score $RSV(d, q)$ et donné par la formule suivante :

$$RSV(d,q) = \frac{\sum_{t \in d \cap q} w_{d,t} \times w_{q,t}}{\sqrt{\sum_{t \in d} w_{d,t}^2} \cdot \sqrt{\sum_{t \in q} w_{q,t}^2}} \tag{2.1}$$

où, $w_{d,t}$ est le poids d'un terme t au sein du document d et $w_{q,t}$ est le poids du terme t au sein de la requête q.

1.2.3 Modèles probabilistes

Le modèle probabiliste est fondé sur la théorie des probabilités. Il suppose l'existence d'une classe de documents pertinents et d'une classe de documents non pertinents pour une requête [Maron and Kuhns, 1960].

Le score de pertinence système est calculé par le rapport entre la probabilité "$p(d|q)$" qu'un document d donné soit pertinent pour une requête q et la probabilité "$p(\overline{d}|q)$" qu'il soit non pertinent pour q. Ces probabilités sont estimées par les probabilités de présence ou d'absence d'un terme de la requête :

- $P(t \in d|\ d$ pertinent) est la probabilité, qu'on notera p, qu'un terme de la requête apparaisse dans un document d sachant que celui ci est pertinent.
- $P(t \in d|\ d$ non pertinent) est la probabilité, qu'on notera q, que ce terme t apparaisse dans un document sachant que celui-ci est non pertinent.

$$RSV(d,q) = \frac{p(d|q)}{p(\overline{d}|q)} = \sum_{t \in q} log \frac{p_t(1-q_t)}{q_t(1-p_t)} \tag{2.2}$$

où, n est le nombre total de termes dans la requête.

Par la suite nous appelons $Pert$ l'ensemble des documents pertinent et $\overline{Pert}$ l'ensemble des documents non pertinents. Nous montrons à l'aide du tableau de contingence (cf. Tableau 2.1) la comptabilisation du principe de présence ou d'absence d'un terme t dans un ensemble de document pertinents $Pert$ et non pertinents $\overline{Pert}$.

	Documents pertinents $Pert$	Documents non pertinents $\overline{Pert}$	Total
Terme t présent	N_t	$df_t - N_t$	df_t
Terme t absent	$\|Pert\| - N_t$	$(N - df_t) - (\|Pert\| - N_t)$	$N - df_t$
Total	$\|Pert\|$	$N - \|Pert\|$	N

Tableau 2.1 – Tableau de contingence

où,

N est le nombre total des documents,

N_t est le nombre de documents pertinents contenant le terme t,

df_t est le nombre des documents contenant le terme t.

Le poids d'un terme dans le calcul du RSV ne prend pas en compte les fréquences d'occurrence du terme t. [Croft and Harper, 1979] ont proposé d'intégrer la fréquence d'occurrence d'un terme dans le calcul du score $RSV(d, q)$. Robertson et *al.*, [Fischer and Reuber, 1977], [Robertson and Walker, 1994], ont ensuite proposé le modèle 2-Poisson qui tient compte de différents aspects relatifs à la fréquence d'occurrence des mots dans les documents et la taille des documents et qui a été à l'origine de la pondération $BM25$, détaillée dans la section 1.4 [Robertson and Walker, 1999].

1.3 Principaux éléments de la fonction de pondération

La pondération d'un terme d'indexation dépend de trois caractéristiques ; la **représentativité** du terme pour un document, son pouvoir **discriminant** pour le document et la **normalisation** par la taille d'un document (et/ou par la taille de la requête dans certains cas).

1.3.1 Représentativité d'un terme pour un document

Plus un document contient d'occurrences d'un terme, plus il est considéré comme abordant une thématique relative à ce terme. Un document d contenant un grand nombre d'occurrences du terme t est considéré comme un document parlant de t et donc plus pertinent pour la requête contenant le terme t. La mesure de représentativité $TF_{d,t}$ d'un terme t pour le document d permettant de refléter l'importance du terme t pour d est basée sur le nombre d'occurrences de t dans d, noté $tf_{d,t}$ (*term frequency*).

1.3.2 Pouvoir discriminant d'un terme dans le corpus

Un terme est discriminant pour un document s'il permet de distinguer le document en question du reste des documents du corpus. Les termes qui apparaissent fréquemment dans tous les documents ne peuvent donc pas être considérés comme discriminants [Salton and Buckley, 1988].

La mesure du pouvoir discriminant d'un terme t est calculée sur la base de la fréquence documentaire inverse appelée IDF_t (inverse document frequency), définie souvent par la formule :

$$IDF_t = \frac{1}{df_t} \tag{2.3}$$

où df_t représente le nombre de documents du corpus dans lesquels le terme apparaît. Plus IDF_t est grand, plus le terme est considéré comme discriminant.

1.3.3 La normalisation par la taille

Les tailles très variables des documents d'un même corpus limitent l'efficacité de la mesure TF pour déterminer l'importance des termes [Harman, 2000]. Le problème majeur est que les termes appartenant aux documents longs ont plus de chance d'apparaître plus fréquemment et d'avoir des poids plus élevés que dans les documents courts. Ainsi, les documents longs se retrouvent avec plus de chance d'être sélectionnés [Singhal et al., 1996].

Pour pallier cet inconvénient, la pondération peut être définie par une combinaison normalisée du pouvoir de discrimination et de représentativité d'un terme :

$$Normalisation(TF_{d,t} \times IDF_t) \tag{2.4}$$

Dans certains cas, il n'y a pas de normalisation appliquée au calcul du poids d'un terme et la pondération d'un terme est simplement égale à la multiplication de mesures de représentativité et du pouvoir discriminant.

1.3.4 Variantes des composantes de pondération

Salton et Buckley proposent plusieurs variantes des composantes de la fonction de pondération $TF.IDF$ et de la normalisation. Voici les principales [Salton and Buckley, 1988] :

Variantes de la composante TF :

- Binaire : $TF_{d,t}$ prend des valeurs binaires.

$$TF_{d,t} = \begin{cases} 1 : si\ t\ apparaît\ au\ moins\ une\ fois\ dans\ le\ document\ d \\ 0 : sinon. \end{cases}$$

- Naturel (*term frequency* "tf") : $TF_{d,t}$ d'un terme dans le document est le nombre d'occurrences $tf_{d,t}$ de ce terme au sein d'un document.

$$TF_{d,t} = tf_{d,t} \tag{2.5}$$

– Logarithmique : la valeur d'un $tf_{d,t}$ brut est donnée par une fonction logarithmique.

$$TF_{d,t} = 1 + log(tf_{d,t}) \tag{2.6}$$

Proposée par Salton et Buckley, cette fonction permet d'atténuer l'écart entre les fréquences d'occurrence des termes au sein d'un document de sorte qu'un document contenant une grande valeur de TF pour un terme de la requête est aussi pertinent qu'un document contenant plusieurs termes de la requête mais avec des valeurs de TF moins élevées pour chaque terme [Salton and Buckley, 1988]. De plus, il y a une différence significative entre 1 et 2 occurrences d'un terme t et donc l'augmentation de son poids doit être importante comme représenté dans la figure 2.3 (différence Δ_1). Par contre il n'y a presque pas de différence entre 1000 et 1001 occurrences et donc l'augmentation de son poids doit être négligeable comme représenté (différence Δ_2) [Robertson and Zaragoza, 2009]. Il s'agit ici du comportement appelé "phénomène de saturation".

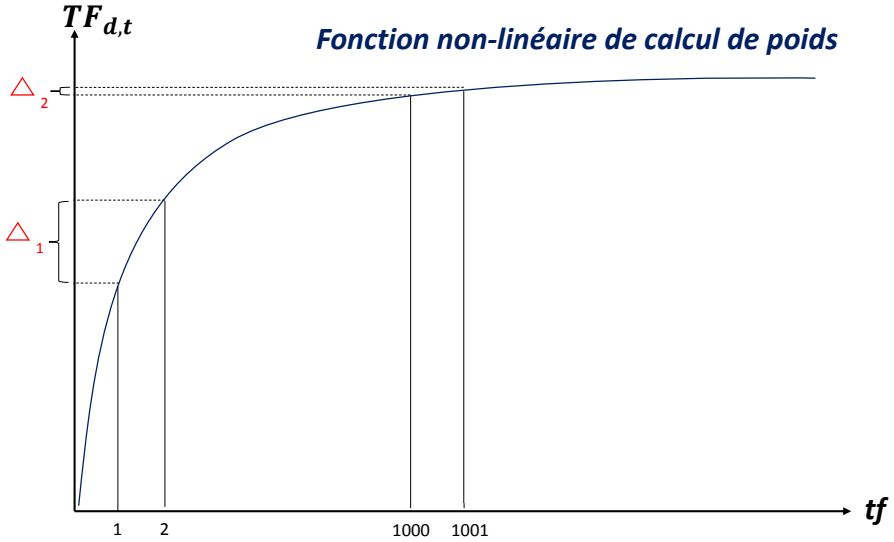

Figure 2.3 – Comportement de saturation

Variantes de la composante IDF :

– Fréquence documentaire inversée : la valeur de l'IDF_t est donnée par la fonction suivante :

$$IDF_t = log\left(\frac{N}{df_t}\right) \tag{2.7}$$

où N est le nombre total des documents du corpus et df_t est le nombre de documents contenant le terme t.

– Probabiliste : la valeur de l'IDF_t est donnée par la fonction probabiliste suivante :

$$IDF_t = log\left(\frac{N - df_t}{df_t}\right) \tag{2.8}$$

Variante de la composante de normalisation :

– TF normalisée : une fonction de normalisation est appliquée à $TF_{d,t}$ de sorte que :

$$Normalisation(TF_{d,t}) = \frac{tf_{d,t}}{MaxTF_{t' \in d}(tf_{d,t'})} \tag{2.9}$$

ou

$$Normalisation(TF_{d,t}) = 0,5 + 0,5\frac{tf_{d,t}}{MaxTF_{t' \in d}(tf_{d,t'})} \tag{2.10}$$

où $MaxTF_{t' \in d}(tf_{d,t'})$ est le nombre maximum d'occurrences d'un terme dans le document d.

– Cosinus : dans ce cas, la normalisation appliquée est une normalisation euclidienne calculée par la formule suivante :

$$Normalisation(TF_{d,t}, IDF_t) = w_{d,t} \times \frac{1}{\sqrt{\sum_{t'}(w_{d,t'})^2}} \tag{2.11}$$

1.4 Pondération Okapi BM25

Okapi $BM25$ est une fonction de pondération basée sur les modèles probabilistes et proposée par Robertson et Spark Jones [Robertson et al., 1996]. Le nom Okapi vient de l'un des premiers systèmes de recherche d'information qui a implémenté

Type de composantes	**Fonctions**
Composantes des TF	
b	1, 0
tf	$tf_{d,t}$
a_n	$0,5 + 0,5\frac{tf_{d,t}}{MaxTF}$
Composantes des IDF	
f	$log(\frac{N}{df_t})$
p	$log(\frac{N-df_t}{df_t})$
Composantes de $Normalisation$	
l	$log(tf_{d,t})$
c	$\frac{1}{\sqrt{\sum_d (w_{d,t})^2}}$

Tableau 2.2 – Composantes de pondération [Salton and Buckley, 1988]

cette fonction à *London's City University*.

$$RSV(d,q) = \sum_{t \in q \cap d} \overbrace{\frac{(k_1+1)tf_{d,t}}{k_1(1-b+b\times\frac{dl}{avgdl})+tf_{d,t}}}^{TF_{d,t}} \times \overbrace{log\frac{N-df_t+0,5}{N-df_t}}^{IDF_t} \times \overbrace{\frac{(k_3+1)tf_{q,t}}{k_3+tf_{q,t}}}^{QTF_{q,t}} \tag{2.12}$$

où :

- q : une requête, contenant un ou plusieurs termes éventuellement pondérés,
- $tf_{d,t}$: fréquence d'apparition du terme t dans le document d,
- $tf_{q,t}$: fréquence d'apparition du terme t dans la requête q,
- N : nombre de documents du corpus,
- df_t : nombre de documents du corpus contenant le terme t,
- k_1, b et k_3 : paramètres de la fonction,
- dl : longueur du document d,
- $avgdl$: la longueur moyenne d'un document dans le corpus.

1.4.1 Représentativité du poids du terme au sein du document : $TF_{d,t}$

$$TF_{d,t} = \frac{(k_1+1)tf_{d,t}}{k_1(1-b+b\times\frac{dl}{avgdl})+tf_{d,t}} \tag{2.13}$$

Cette partie représente le poids du terme t dans le document d, en nombre d'oc-

currences ($tf_{d,t}$) de t, normalisé par la proportion de la taille du document et la taille moyenne des documents du corpus. La normalisation par taille est contrôlée par le paramètre b. En plus de la normalisation par la taille, la formule BM25 fournit le moyen de contrôler l'effet de la saturation au niveau des occurrences des termes du document $tf_{d,t}$ grâce au paramètre k_1 (cf. figure 2.3). Ce comportement de contrôle de saturation vise à donner moins d'importance à chaque nouvelle occurrence d'un terme dans le même document.

1.4.2 Pouvoir discriminant du terme : IDF_t

$$IDF_t = log\left(\frac{N - df_t + 0,5}{N - df_t}\right) \tag{2.14}$$

Cette partie de la formule $BM25$ permet de calculer le pouvoir discriminant du terme t pour un document à travers le corpus des documents avec un IDF_t normalisé donné par une fonction logarithmique.

1.4.3 Poids du terme au sein de la requête : $QTF_{d,t}$

$$QTF_{q,t} = \frac{(k_3 + 1)tf_{q,t}}{k_3 + tf_{q,t}} \tag{2.15}$$

Cette partie permet de calculer un poids du terme t dans la requête q. De même que pour les occurrences des termes au sein du document, la saturation des occurrences des termes au sein de la requête peut être contrôlée par le paramètre k_3 de la formule BM25.

1.5 Modèle $BM25F$

Le modèle $BM25F$, proposé par Robertson et *al.*, est un modèle de pondération utilisé dans le cas des documents structurés composés de différents champs d'information (ex : titre, résumé, contenu, etc.) où F (*Field*) fait référence aux différents champs au sein du document [Robertson et al., 2004].
Nous présentons dans cette section, le modèle BM25F proposé par Zaragoza et *al* [Zaragoza et al., 2004], que nous avons utilisé dans l'une de nos contributions, et qui est une version étendue du BM25F proposé par Robertson et *al* [Robertson et al., 2004].

$$BM25F(d,q) = TF_{d,t} \times IDF_t \times QTF_{q,t} \tag{2.16}$$

$$TF_{d,t} = \frac{TF_{F_{d,t}}}{k_1 + TF_{F_{d,t}}} \tag{2.17}$$

tel que :

$$TF_{F_{d,t}} = \sum_{f \in d} w_f \times \overline{tf_{f_{d,t}}} \tag{2.18}$$

avec :

- $TF_{F_{d,t}}$ est la fréquence d'occurrence du terme t dans l'ensemble des champs (F) du document d,
- IDF_t et $QTF_{q,t}$ représentent respectivement IDF_t et $QTF_{q,t}$ classiques dans le modèle BM25 [Robertson et al., 1996],
- k_1 représente le paramètre classique k_1 dans le modèle BM25 [Robertson et al., 1996],
- w_f représente le poids attribué à chaque champ f du document,
- $tf_{f_{d,t}}$ est la fréquence d'occurrence du terme t dans le champ f du document d,
- $\overline{tf_{f_{d,t}}}$ est la version normalisée du $tf_{f_{d,t}}$.

$$\overline{tf_{f_{d,t}}} = \frac{tf_{f,t}}{(1 - b_f) + b_f \frac{fl}{avgfl}} \tag{2.19}$$

avec :

- b_f correspond au paramètre b du modèle $BM25$ pour le champ f du document d,
- fl et $avgfl$ représentent respectivement la taille du champ f du document d et la taille moyenne des champs.

1.6 Axiomes de RI

Les fonctions de correspondance et de calcul de score sont un élément clé des modèles de RI. Leurs caractéristiques, notamment celles concernant les distributions des fréquences de termes, jouent un rôle important dans le comportement d'un modèle de RI.

Fang et al. proposent une liste de contraintes heuristiques permettant de vérifier

qu'un modèle de RI est bien fondé [Fang et al., 2004].
Afin de vérifier la compatibilité de leur modèle de RI avec ces contraintes heuristiques, Clinchant et Gaussier présentent une étude analytique permettant d'explorer les liens entre ces contraintes heuristiques et les différentes caractéristiques de la fonction de pondération proposée dans leur modèle de RI fondé sur les lois de probabilité [Clinchant and Gaussier, 2010].
Nous présentons brièvement les principaux axiomes de ces contraintes heuristiques pour expliquer l'impact de la variation des éléments de la fonction de pondération (distribution des termes, taille des documents) sur le calcul de score.
Contraintes sur les TF :

1. Axiome TF_1 : étant donné un terme t dans la requête q et un document d
   $tf_{d,t} > 0 \Longrightarrow RSV(d,q) > 0$.
   Si le terme t de la requête apparaît dans le document d au moins une fois, le score de correspondance entre la requête q et le document d est non nul.
2. Axiome TF_2 : étant donné deux termes t et t' dans la requête q et deux documents d et d'.
   Toutes choses étant égales, par ailleurs $df_t = df_{t'}$ en particulier $dl = dl'$, on doit avoir :
   - Si la somme des fréquences d'apparition des deux termes de la requête t et t' dans le document d est égale à celle dans d', alors les scores de correspondance des deux documents d et d' pour la requête q sont égaux.
     $(tf_{d,t} + tf_{d,t'}) = (tf_{d',t} + tf_{d',t'}) \Longrightarrow RSV(d,q) = RSV(d',q)$.
   - Si la somme des fréquences d'apparition des deux termes de la requête t et t' dans le document d est plus grande que dans d', alors le score de d pour la requête q est plus grand que celui de d'.
     $(tf_{d,t} + tf_{d,t'}) > (tf_{d',t} + tf_{d',t'}) \Longrightarrow RSV(d,q) > RSV(d',q)$.
3. Axiome TF_3 : étant donné un terme t dans la requête q et trois documents d, d' et d''.
   - Si le document d'' a une plus grande fréquence d'occurrence du terme t de la requête q que celle du terme t dans le document d', alors d'' a un plus grand score que le document d' pour q. De même que d' ayant une plus grande fréquence d'occurrence de t par rapport à d aura un score plus grand que celui de d. Notons surtout qu'en vertu du phéno-

mène de saturation, pour une même variation du nombre d'occurrences du terme dans une requête, l'augmentation du score sera moindre lorsque les nombres d'occurrences sont élevés (par exemple pour une variation de 1000 à 1001) que lorsqu'il sont petits (par exemple de 1 à 2) [Fang et al., 2004].
$dl = dl' = dl''$ et $tf_{d'',t} - tf_{d',t} = 1$ et $tf_{d',t} - tf_{d,t} = 1 \Longrightarrow RSV(d', q) - RSV(d, q) > RSV(d'', q) - RSV(d', q)$

Contraintes sur les longueurs normalisées NL :

1. Axiome NL_1 : étant donné deux termes t et t' dans la requête q et deux documents d et d'.
   - Si la taille du document d' est plus grande que la taille du document d ($dl' > dl$), et si la somme des fréquences d'apparition des deux termes de la requête t et t' dans le document d est égale à celle dans d', le score de correspondance du document d pour la requête q doit être plus important que celui de d'.
     $dl' > dl$ et $(tf_{d,t} + tf_{d,t'}) = (tf_{d',t} + tf_{d',t'}) \Longrightarrow RSV(d, q) \geq RSV(d', q)$.

2. Axiome $NLTF_2$: étant donné un terme t dans la requête q et deux documents d et d'.

   En considérant que la requête comporte une seule occurrence de chaque terme, le score de correspondance entre la requête et le document est souvent donné par l'équation suivante :

   $$RSV(d, q) = \sum_{t \in q \cap d} w_{d,t} \tag{2.20}$$

   Si le document d' est k fois plus grand que le document d ($\forall k > 1$) et si la fréquence d'occurrence $tf_{d',t}$ du terme t dans le document d' est k fois plus grande que sa fréquence d'apparition ($tf_{d,t}$) dans d alors le poids du terme t dans d' est plus grand que celui dans d. Ainsi, le score du document d' est plus important que celui du document d pour la requête q.
   - $dl' = k \times dl$ et $tf_{d',t} = k \times tf_{d,t} \Longrightarrow w_{d',t} > w_{d,t}$
   - $w_{d',t} > w_{d,t} \Longrightarrow RSV(d', q) > RSV(d, q)$.

Contraintes sur les IDF des mots qui apparaissent plus souvent :

1. Axiome IDF_1 : étant donné deux termes t et t' dans la requête q et deux documents d et d' contenant respectivement t et t' ($t' \notin d$ et $t \notin d'$).

   Les documents ayant des termes qui ont un IDF_t plus élevé en plus d'une fréquence élevée (cf. Axiome TF_2) obtiennent un score élevé.

   En ayant les conditions de l'axiome TF_2 vérifiées : $dl = dl'$:

   – $idf_{t'} \geq idf_t$ et $tf_{d',t'} = tf_{d,t} \Longrightarrow RSV(d', q) \geq RSV(d, q)$

1.7 Évaluation en RI

L'évaluation en RI est une étape permettant de mesurer la capacité des systèmes de RI à satisfaire les besoins d'information des utilisateurs. Cette évaluation repose sur la notion de pertinence des résultats.
La notion de pertinence est souvent difficile à définir et qualifier. Le document est pertinent s'il répond au besoin d'information déclaré par l'utilisateur. Cependant, la perception de l'utilisateur ne coïncide pas toujours avec la notion de qualité des systèmes [Manning et al., 2008].
On considère deux principaux types de pertinence :
– la pertinence utilisateur : l'utilisateur lui même juge le document pertinent en fonction de son besoin d'information.
– la pertinence système : le SRI juge le document pertinent pour une requête sur la base de la fonction de pertinence ($RSV(d, q)$).
La qualité d'un modèle de RI est mesurée en comparant les résultats retournés par le système (pertinence système) avec les attentes de l'utilisateur (pertinence utilisateur) (cf. figure 2.4).
La pertinence utilisateur doit être la plus proche possible de la pertinence système.

1.7.1 Collection de test en recherche d'information

Pour évaluer les systèmes de RI, des collections de test sont requises. En s'inspirant de la citation de Voorhees et Buckley "*A test collection defines the common task that allows the effectiveness of different retrieval mechanisms to be directly compared*" [Voorhees and Buckley, 2002], nous définissons une collection de test comme étant l'ensemble de données et d'informations conçues pour évaluer la qualité d'un SRI en terme de rapprochement de la pertinence système de la pertinence utilisateur.

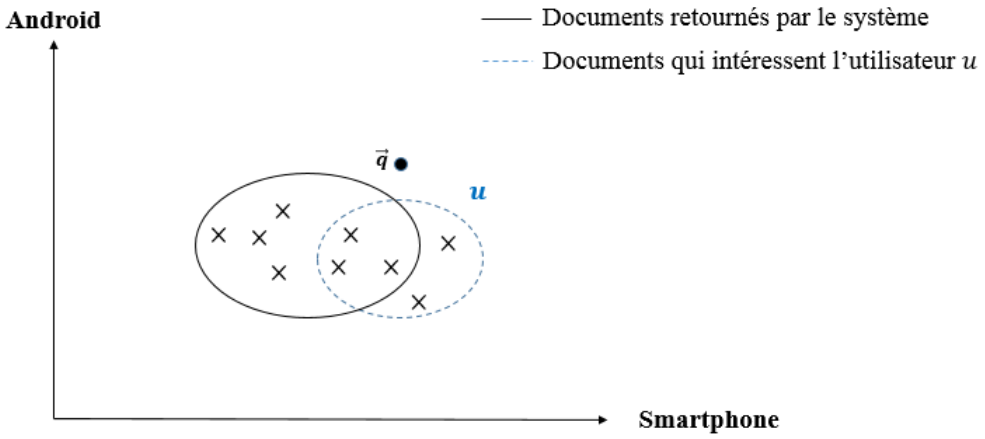

Figure 2.4 – Évaluation en recherche d'information

Une collection de test de RI (cf. figure 2.5) est composée de :

- **Un corpus de documents** D : le corpus de documents est une liste de documents disponibles sur lesquels la recherche d'information est effectuée.
- **Un ensemble de besoins d'informations** Q : les requêtes permettent aux utilisateurs de formuler leurs besoins d'information par une liste d'un ou plusieurs mots clés.
- **Un ensemble de jugements de pertinence** $Qrels_q$: le système dispose des jugements de pertinence utilisateur. Ils permettent d'indiquer pour chaque document du corpus pour chaque besoin d'information s'il est pertinent, en précisant éventuellement son degré de pertinence. Un jugement de pertinence est un couple (document, requête) ou un triplet (document, requête, degré)

Un des premiers paradigmes dans les approches d'évaluation est le paradigme nommé *Cranfield* [Cleverdon, 1997], dans lequel l'évaluation se base sur la comparaison des réponses retournées par le système pour une requête avec les réponses dites "idéales" qui ont été identifiées manuellement par des experts dans le domaine de recherche.

Dans les collections Cleverdon, l'objectif principal était de tester l'efficacité des bibliothécaires à localiser des documents indexés par plusieurs systèmes afin de répondre aux demandes des utilisateurs de la bibliothèque pour retrouver des références bibliographiques.

Dans Cleverdon 1997, il n'est pas suffisant que les groupes évaluent leurs propres systèmes. Ainsi, il a proposé quatre compétitions qui indexaient avec différentes

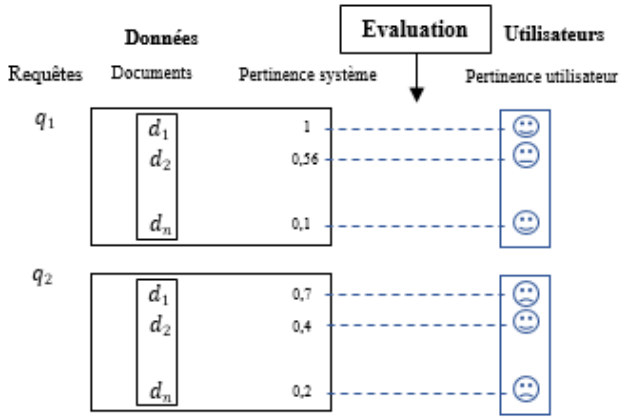

Figure 2.5 – Collection de test et évaluation en recherche d'information

approches sur une collection de 18.000 articles.
Les jugements de pertinence étaient exhaustifs et obtenus manuellement par les juges experts. Par la suite, des systèmes de *pooling* ont été proposés, où les jugement de pertinence sont attribués seulement à un sous ensemble de documents de la collection correspondant aux top k-documents retournés par les différents systèmes de RI participant à l'évaluation.

Collections de test classiques :

Il existe de nombreuses collections de test en RI utilisées dans les campagnes d'évaluation. Nous citons les collections de tests "*Cranfield*" et les principales compagnes d'évaluation qui se basent sur le même principe que celui des collections "*Cranfield*", à savoir, les compagnes de TREC, INEX, CLEF, etc.
Dans ces collections, les jugements de pertinence sont construits manuellement de manière collaborative par les utilisateurs des systèmes. On peut citer quelques collections de test standard [Manning et al., 2008] :

– La collection Cranfield [3] : La pionnière des collections de test qui permet d'avoir une mesure quantitative précise de l'efficacité de RI. Collectée au Royaume Uni, elle contient 1.398 résumés d'articles, un ensemble de 225 requêtes et des jugements de pertinence binaire exhaustifs de toutes les paires "requêtes - documents".
– La compétition TREC (*Texte REtrieval Conference*) [4] : L'institut national des

3. `http://ir.dcs.gla.ac.uk/resources/test_collections/cran/`
4. `http://trec.nist.gov/`

standards et technologies (NIST) a lancé de grandes séries de construction de jeux de données d'évaluation depuis 1992. Ces collections comprenaient initialement 1,89 millions de documents et des jugements de pertinence pour 450 besoins d'information.

- La compétition NTCIR (*NII Collection for Information Retrieval*) [5] : Le projet NTCIR a permis la construction de différentes collections de test avec des tailles similaires aux collections TREC se focalisant sur le langues asiatiques et le *cross-language information retrieval*, où les requêtes sont établies dans une collection de documents contenant une ou plusieurs langues.
- La compétition CLEF (*Cross Language Evaluation Forum*) [6] : Cette série d'évaluation se focalise dans des langues européennes ainsi que la recherche d'information en multi-langues (*cross-language information retrieval*).

Collections de test générées automatiquement :

Voorhees et Harman [Voorhees and Harman, 2005] pointent deux limitations importantes dans la construction de collections de test de RI classique :

- La première est liée à la pertinence des données et la difficulté de déterminer exhaustivement la pertinence de l'utilisateur par des jugements manuels d'un expert (juge), sachant que souvent les juges ne peuvent pas estimer pertinemment les attentes de l'utilisateur, comme montré dans les études analytiques faites dans [Chapelle and Zhang, 2009] et dans [Agrawal et al., 2009].
- La seconde est principalement liée aux moyens mis en oeuvre dans le processus de construction des collections de test, qui s'avèrent coûteux [Allan et al., 2008], [Carterette et al., 2008], [Mccreadie et al., 2013].

Pour régler le problème de coût engendré par la génération manuelle et collaborative des collections de test en RI et l'absence des jugements de pertinence dans la plupart des cas, certains travaux ont opté pour la construction automatique de collections de test en proposant des approches permettant de générer automatiquement des requêtes et des jugements de pertinence.

De nombreuses approches supposent qu'il est possible de prédire la pertinence de l'utilisateur et ses attentes en se servant de son comportement sur le Web à travers les clics, les temps de réponses ou le temps passé sur une page ([Joachims, 2002a], [Raiber and Kurland, 2013]), les annotations et commentaires des utili-

5. `http://research.nii.ac.jp/ntcir/index-en.html`
6. `http://clef.isti.cnr.it/`

sateurs sur le Web, etc. ([Vallet et al., 2010], [Xu et al., 2008], [Schenkel et al., 2008], [Carmel and Yom-Tov, 2010], [Raiber and Kurland, 2013]). Ce comportement de l'utilisateur sur le web peut fournir des informations à propos de la pertinence de l'utilisateur en termes des documents (ressources) que l'utilisateur souhaiterait avoir dans la liste des résultats retournés. Les travaux de recherches connexes estiment que l'utilisateur pourrait être intéressé par des documents ou des ressources similaires aux documents et liens qu'il parcourt.

1.7.2 Mesures d'évaluation

Afin d'évaluer la pertinence du système en terme de documents retournés pour un besoin d'information spécifique, des mesures d'évaluation standard peuvent être appliquées, telles que le Rappel et la Précision, la moyenne des précisions moyennes (MAP), la précision $P[0.x]$ pour un taux de rappel x donné ou la *F-mesure* :

Le Rappel, s'interprétant comme la capacité à retrouver **tous** les documents pertinents, est le nombre de documents pertinents retrouvés par rapport au nombre de documents existants (cf. équation 2.21, figure 2.6).

$$Rappel = \frac{|PR|}{|P|} = \frac{nombre\ \ de\ \ document\ \ pertinents\ \ retrouvés}{nombre\ \ de\ \ documents\ \ pertinents} \tag{2.21}$$

avec :

– PR : l'ensemble des documents pertinents retrouvés,

– P : l'ensemble des documents pertinents pour la requête q,

La Précision représentant la capacité à ne retrouver **que** les documents pertinents, est définie comme le nombre de documents pertinents retrouvés par rapport au nombre de documents retrouvés (cf. figure 2.6).

$$Précision = \frac{|PR|}{|PR \bigcup NPR|} = \frac{nombre\ \ de\ \ documents\ \ pertinents\ \ retrouvés}{nombre\ \ de\ \ documents\ \ retrouvés} \tag{2.22}$$

avec :

– R : l'ensemble des documents retrouvés,

– NPR : l'ensemble des documents non pertinents retrouvés,

L'avantage des deux mesures (précision et rappel) est qu'elles sont complémen-

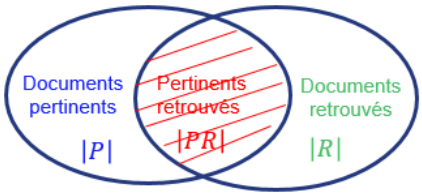

Figure 2.6 – *Rappel - Précision*

taires et que chacune peut être privilégiée selon les circonstances : un utilisateur effectuant une recherche sur le Web souhaite souvent que les premiers documents retournés soient pertinents (haute *précision*) et il accorde moins d'intérêt au fait d'obtenir le plus grand nombre possible de documents pertinents (faible rappel). A l'inverse, certains professionnels comme par exemple dans le domaine juridique, sont plus intéressés par avoir le plus grand nombre possible de documents pertinents retournés (un grand *rappel*) sans accorder autant d'importance au taux de *précision* [Manning et al., 2008].

La mesure appelée *F-mesure* permet de combiner la précision et le rappel en une seule et unique mesure.

F-mesure : la F-mesure est une moyenne harmonique pondérée de la *précision* et du *rappel* :

$$\begin{aligned} F &= \frac{1}{\alpha\frac{1}{Precision} + (1-\alpha)\frac{1}{Rappel}} \\ &= \frac{(\beta^2+1)Precision \times Rappel}{\beta^2(Precision + Rappel)} \end{aligned}$$

où $\beta^2 = \frac{1-\alpha}{\alpha}$, $\alpha \in [0,1]$ et donc $\beta^2 \in [0,\infty[$

La F-mesure équilibrée appelée généralement *F1-mesure* permet de pondérer de façon équilibrée la *précision* et le *rappel* où les valeurs de $\alpha = \frac{1}{2}$ et $\beta = 1$ [Manning et al., 2008].

$$F_{\beta=1} = \frac{2 \times Precision \times Rappel}{Precision + Rappel} \tag{2.23}$$

Souvent, l'évaluation globale d'un SRI fait appel à la courbe de *précision - rappel*

(cf. figure 2.7), à la *précision*, et à la MAP (moyenne des précisions moyennes), que nous utilisons.

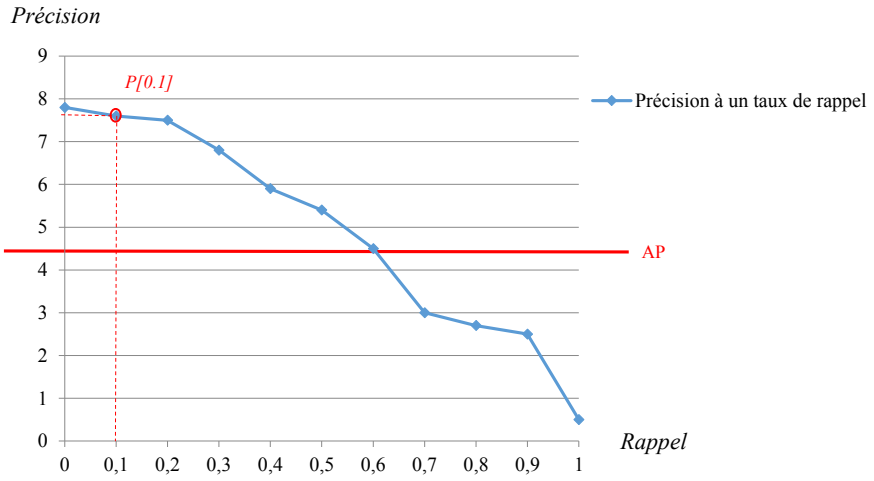

Figure 2.7 – Courbe de *Rappel* - *Précision*

La précision à un rang donné ***(***$P@k$***)*** : c'est la précision pour les k premiers documents retournés.

La précision à un taux de rappel ***(***$P[0.x]$***)*** : c'est la précision interpollée obtenue quand un taux de rappel à x est atteint. Dans l'exemple de la figure 2.7, pour un taux de rappel de 10% $P[0.1] = 0,78$.

La précision moyenne ***(***AP_l***)*** : la précision moyenne pour AP_l est calculée pour chaque requête q_l. Il s'agit de la moyenne des *Précisions* à chaque rang de document pertinent pour cette requête. Elle est donnée par la formule 2.24.

$$AP_l = \frac{\sum_{k=1,K}(rel(k).P@k)}{\sum_{k=1,K} rel(k)} \tag{2.24}$$

Avec K le nombre de documents retrouvés, $P@k$: la précision au rang k et $rel(k)$ calculé comme suit : $rel(k) = \begin{cases} 1 : si \ \ le \ \ k^{ieme} \ \ document \ \ est \ \ pertinent \\ 0 : sinon. \end{cases}$

La moyenne des précisions moyennes (Mean Average Precision "MAP") : La moyenne des *précisions* moyennes (***MAP***) pour l'ensemble des requêtes est calculée à l'aide des précisions moyennes AP_l des requêtes, afin de mesurer la performance

globale du système ([Manning et al., 2008]) comme présenté dans l'équation 2.25.

$$MAP = \frac{1}{|Q|} \sum_{l=1,|Q|} AP_l \tag{2.25}$$

La précision moyenne à un taux de rappel est une précision moyenne calculée à 11 points de rappel (0% ... 100%) par un pas de 10%. Elle consiste à moyenner les 11 précisions obtenues pour les 11 seuils de rappel.

2 Recherche sociale personnalisée d'information

"*Both the user's information needs and his strategies for satisfying them are influenced by the socio-cultural environment, since they arise in social situations*" [Wilson, 1981]. Ainsi, la recherche d'information impliquant des utilisateurs au sein des réseaux sociaux doit tenir compte des informations sociales issues de l'environnement social dans lequel se trouve l'utilisateur afin de mieux satisfaire ses besoins d'information.

La personnalisation en recherche d'information permet d'adapter les processus de RI afin de retourner des résultats appropriés aux utilisateurs selon leurs centres d'intérêts.

2.1 Réseaux sociaux

Les réseaux sociaux sont un espace dans lequel les internautes interagissent (publient, partagent, annotent, commentent, etc.) avec le contenu du Web [Fischer and Reuber, 2010]. Il peut s'agir d'images (Flickr : 6 miliards de photos [7]), de ressources (Twitter : plus de 500 millions d'utilisateurs, Facebook : plus d'un miliard d'utilisateurs [8], Delicious), ou encore d'informations professionnelles (LinkedIn : 175 millions de membres [9]). Les réseaux sociaux représentent aussi un moyen de communication et d'échange efficace en permettant aux utilisateurs de rentrer en contact avec des collègues, amis, co-auteurs, etc.

Avec l'émergence des réseaux sociaux, l'utilisateur d'un système de RI n'est pas

7. http ://fr.slideshare.net/WaveLab/10-social-networks-and-a-bunch-of-stats
8. http ://www.zdnet.fr/actualites/facebook-un-milliard-d-utilisateurs-actifs-dans-le-monde-39783232.htm
9. http ://press.linkedin.com/about

considéré comme un acteur isolé. Son besoin d'information est vu au sein d'un réseau dans un contexte social décrit par un contenu social (tags, annotations, citations, tweets, statuts, "j'aime", etc.) et des relations sociales (amis, co-auteurs, suiveurs, etc.).

2.2 Exploitation des informations sociales pour la RI

L'exploitation des réseaux sociaux a permis d'améliorer la RI de différentes façons en rajoutant de nouvelles informations supplémentaires sur les ressources (documents, utilisateurs, etc.) :

– L'annotation d'un document par plusieurs utilisateurs intéressants et renommés dans le réseau social peut signaler un document populaire, fiable, etc. ([Bao et al., 2007], [Bouadjenek et al., 2013], [Soulier et al., 2012]).

– Un utilisateur cité ou référencé peut représenter généralement une source fiable, un individu expert ou un utilisateur populaire et influant dans son entourage, etc. ([Schenkel et al., 2008],[Bao et al., 2007]).

– Un utilisateur est plus intéressé par des documents qui lui sont fournis ou suggérés par son entourage (groupe d'individus connus par l'utilisateur) que par des documents venant d'individus inconnus (notion de confiance [Kirsch, 2005]).

– En connaissant le profil de l'utilisateur (initiateur de requête) et son domaine d'intérêt, à partir des informations qui le caractérisent au sein de son réseau social, le système de recherche d'information serait capable de retourner des résultats de recherche répondant le mieux aux attentes de l'initiateur de requête.

2.3 Problématiques de recherche sociale d'information

Dans la littérature, la recherche sociale personnalisée d'information RPSI est vue comme un sous domaine de la RI qui assiste l'utilisateur d'un réseau social dans sa recherche d'information en exploitant l'expertise et l'expérience des autres utilisateurs dans un contexte où les utilisateurs ayant déjà trouvé des informations pertinentes auront la volonté de les partager avec leur entourage [Goh and Foo, 2008].

La RSPI peut être définie comme étant l'incorporation des annotations et relations sociales, issues des réseaux sociaux (informations sociales), dans le processus de recherche d'information [Kirsch, 2005].

Le modèle de recherche sociale d'information est centré autour des utilisateurs dans leurs contextes sociaux.

Dans la pratique, deux utilisateurs peuvent formuler la même requête au SRI et juger différemment la pertinence des documents retournés. Le principal problème est la pertinence de l'information retournée relativement au contexte de l'utilisateur avec des besoins d'informations spécifiques.

Une des idées de base est d'apprendre les centres d'intérêt de l'utilisateur à travers son comportement sur le Web (ses annotations, ses préférences en terme de documents annotés, etc.) et de construire ensuite un profil utilisateur reflétant ses centres d'intérêt.

Si comme dans le chapitre 1 on considère le cas des deux utilisateurs *Alice* et *Bob* avec des centres d'intérêt différents qui formulent la même requête mais ayant des attentes différentes selon leurs centres d'intérêt [Harter, 1992], [Mizzaro, 1997], le système de RSPI doit pouvoir tenir compte des centres d'intérêt de chaque utilisateur pour satisfaire son besoin d'information exprimé par sa requête et retourner des listes de documents potentiellement différentes pour chaque utilisateur selon son profil.

La principale problématique est de savoir d'une part, comment les modèles de RI doivent s'adapter pour prendre en compte ces nouvelles informations, d'autre part, comment évaluer des modèles de RSPI tenant compte de ces informations sociales à propos des utilisateurs en l'absence de collections de test et de compétitions dédiées.

L'objectif ensuite est de tenir compte du profil de l'utilisateur lors de sa recherche d'information pour mieux répondre à ses besoins d'information en permettant au système de RSPI de lui retourner les documents pertinents selon son profil et ses centres d'intérêts.

Les différentes problématiques liées à ce sujet sont les suivantes :

– **Exploitation des informations sociales**, quelles sont les informations dans les réseaux sociaux qui peuvent être exploitées pour représenter le profil social de l'utilisateur au sein de son réseau social ? Faut-il pondérer et normaliser les termes composant le profil de l'utilisateur ?

– **Intégration du profil de l'utilisateur**, afin de permettre aux systèmes de RSPI de tenir compte de ce profil lors des prochaines recherches de cet utilisateur : à quel niveau du modèle de RSPI faut-il intégrer ces informations ?
Comment combiner ces informations représentant le profil social de l'utilisateur avec les informations des documents ?

– **Évaluation des modèles de RSPI centré utilisateur** : comment peut-on évaluer les modèles de RSPI, avec des utilisateurs ayant des attentes différentes, pour une même requête ?

2.4 Modèle de recherche sociale personnalisée d'information

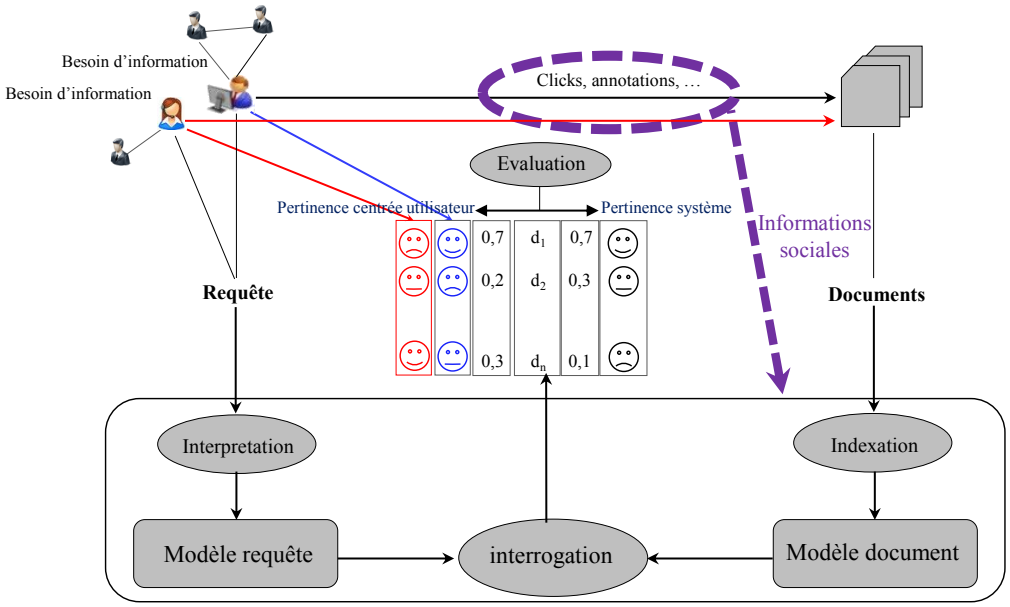

Figure 2.8 – Modèle de recherche sociale d'information

Dans la figure 2.8 les utilisateurs avec un besoin d'information et les documents sont caractérisés par des informations sociales (IS), qui peuvent être à titre d'exemple, le contenu informationnel social des utilisateurs décrit par des annotations. La figure 2.8 montre des informations sociales que le modèle de RSPI peut exploiter. Ces informations sociales apportent de l'information sur les utilisateurs et sur les documents. Il s'agit d'exploiter ces IS au niveau des différentes étapes du processus de RI : indexation, interrogation.

2.5 Évaluation en recherche sociale d'information

La figure 2.9 montre que dans le cas de deux utilisateurs avec des préférences et centres d'intérêt différents, et donc des besoins d'information différents mais for-

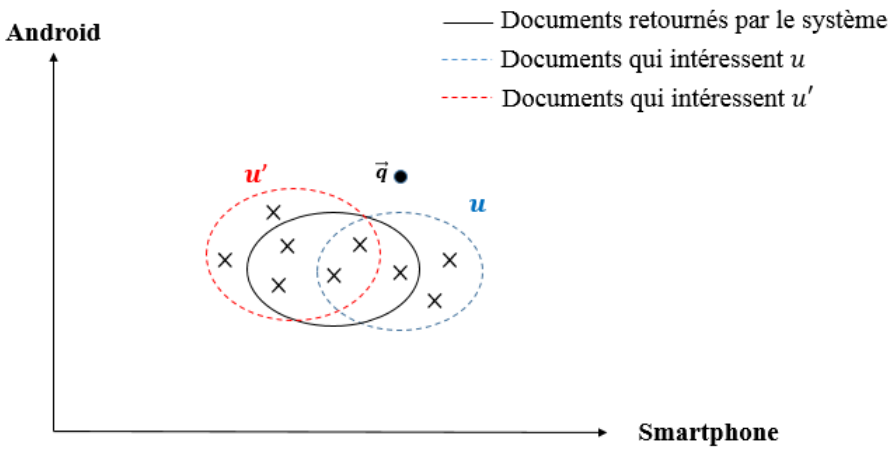

Figure 2.9 – Évaluation en recherche d'information

mulant la même requête au système de RI, le système de RI classique retourne une même liste de résultats pour les deux utilisateurs u et u'.

L'évaluation en RSPI doit considérer chaque requête pour chaque utilisateur comme étant une requête différente représentant un besoin d'information différent et des listes de jugements de pertinence, pour une requête, qui soient potentiellement différentes pour chaque requête et pour chaque utilisateur. Ainsi, un système de RSPI doit pouvoir retourner des résultats personnalisés pour chaque utilisateur tenant compte de ces informations sociales afin d'obtenir une meilleure correspondance entre la pertinence système et la pertinence centrée utilisateur (cf. figure 2.10).

3 Conclusion

Dans ce chapitre, nous avons présenté des généralités sur le domaine de la recherche d'information, à savoir, les grands composants d'un modèle de recherche d'informations, les principaux modèles de RI, et nous avons abordé la question de leur évaluation, notamment au travers des collections de test et des mesures de performance en RI. Nous avons vu ensuite que les réseaux sociaux ont conduit à l'émergence d'une nouvelle thématique : une recherche sociale et/ou personnalisée d'information (RSI / RSPI), qui permettent d'exploiter les informations sociales à propos de l'utilisateur issues des réseaux sociaux dans l'objectif d'améliorer les résul-

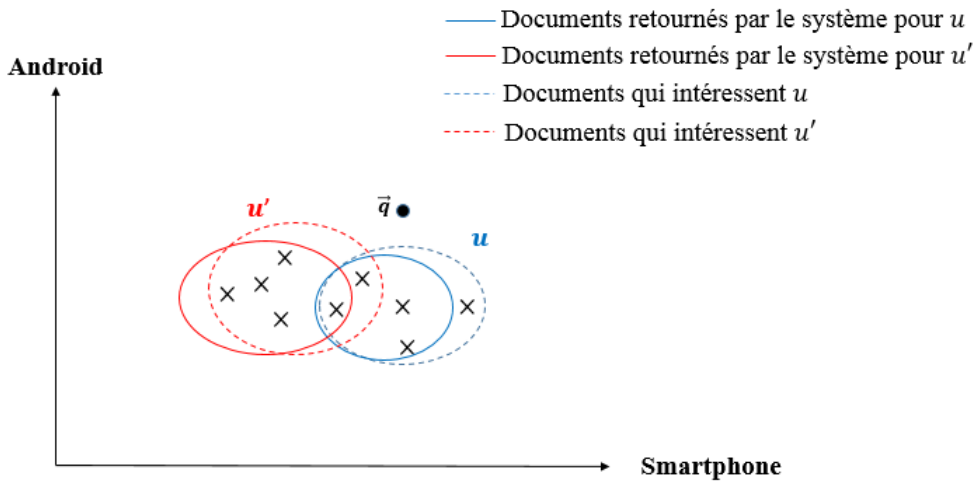

Figure 2.10 – Évaluation en recherche sociale d'information

tats retournés à l'utilisateur compte tenu de ses centres d'intérêt et ses préférences. Nous avons recensé les principales problématiques de la RSPI liées au modèle de RSPI intégrant des informations sociales, et à l'évaluation de ces modèles.

Chapitre 3

État de l'art

1 Introduction

Un des moteurs de recherche sociale sur le Web les plus connus est Social-Search [1], qui permet d'effectuer une recherche au sein des réseaux sociaux reliant un ensemble d'utilisateurs. Ce moteur étend la plate-forme d'annotation social *folkd.com* pour permettre une recherche d'information en exploitant des informations sociales des utilisateurs aux sein des réseaux sociaux.

Dans ce chapitre, nous présentons un aperçu des travaux de l'état de l'art dans le domaine de la recherche sociale et/ou personnalisée d'information (RSPI).

Nous présentons, dans une première partie, les travaux d'état de l'art consacré à l'identification et à l'intégration des informations sociales au sein des modèle de RI à différents niveaux. Nous décrivons ensuite quelques travaux connexes portant sur l'utilisation des informations sociales dans les systèmes de recommandation et de filtrage collaboratif.

Beaucoup de recherches ont été consacrées aux réseaux sociaux et aux données qu'ils engendrent [Newman, 2003], [Stattner and Collard, 2012], notamment l'analyse des annotations sociales, leur sémantique, leur type, leur distribution [Halpin et al., 2007], [Bischoff et al., 2008], [Carman et al., 2009], [Heymann et al., 2008].

Certains travaux se focalisent sur l'étude des caractéristiques des annotations sociales dans les pages Web. Il mettent l'accent sur l'importance de ces annotations pour les page Web [Bischoff et al., 2008] et introduisent des mesures telles que la popularité d'une page Web évaluée à travers le nombre de fois où elle a été annotée [Heymann et al., 2008]. D'autres travaux portent sur l'étude de la structure des réseaux sociaux et le voisinage social de l'utilisateur au sein du réseau [Das et al., 2008], [Kashyap et al., 2012] dans l'objectif d'identifier les préférences de l'utilisateur, de personnaliser la recherche sur le Web [Schenkel et al., 2008], [Xie

1. `http://www.social-search.com`

et al., 2012], [Kashyap et al., 2012] ou de rapprocher les centres d'intérêt partagés par des utilisateurs du réseau [Das et al., 2008].

Les médias sociaux sont souvent représentés par un graphe d'entités sociales : utilisateurs, ressources, contenu social (annotations, commentaires, tweets, notes) et des liens sociaux entre ces entités (liens entre les utilisateurs, entre les ressources et entre les utilisateurs et les ressources), en distinguant différents types de liens, entre les entités du graphe social [Xu et al., 2008], [Schenkel et al., 2008], [Kashyap et al., 2012], etc.

Les informations sociales issues des médias sociaux ont souvent été exploités pour permettre une personnalisation des systèmes et processus de recherche d'information. Ces informations sociales peuvent être vues comme un indicateur d'une importance sociale, généralement estimée à travers un poids social attribué aux entités du graphe social, comme elles peuvent être exploitées pour affiner la description sociale de ces entités (à la fois la ressource et l'utilisateur).

Nous présentons dans la suite, l'extraction des informations sociales selon deux angles : comme indicateur d'importance sociale ou comme profil social.

Nous allons aborder aussi les réseaux sociaux et les systèmes de recommandation et de filtrage collaboratif pour élargir l'état de l'art à des domaines connexes dans lesquels sont proposées des approches qui peuvent nous être utiles pour l'exploitation des informations sociales.

Une seconde partie traitera les travaux d'état de l'art liés à l'exploitation des informations sociales pour la construction des éléments d'une collection de test de RSPI, notamment la construction des requêtes et jugements de pertinence.

2 Indicateur d'importance sociale

Plusieurs travaux proposent de bénéficier des informations sociales issues des médias sociaux pour estimer l'importance sociale des entités au sein des médias sociaux. Les premiers d'entre eux proposent d'adapter l'algorithme du *PageRank* [Brin and Page, 1998] au cas des réseaux sociaux ([Hotho et al., 2006], [Bao et al., 2007], [Kashyap et al., 2012]). Le principe de base dans le *PageRank* est qu'une page est considérée d'autant plus importante qu'elle possède de nombreux liens (ou pages) entrants et que ces liens (pages) sont importants. En plus des liens entrants, l'adaptation du *PageRank* au cas social permet de tenir compte des données sociales

(annotations, commentaires, avis, etc.) associées aux pages et des activités sociales (d'annotation, de référence, d'interaction, etc.) des utilisateurs au sein des réseaux sociaux.

Dans le cas des *Folksonomies*, une entité du graphe social est considérée importante si elle est annotée avec des termes importants employés par des utilisateurs importants ([Hotho et al., 2006], [Bao et al., 2007]).

Hotho et *al.* proposent un algorithme appelé *FolkRank* qui est une adaptation du *PageRank* au principe du cas social [Hotho et al., 2006]. Dans la même catégorie, Bao et *al.* proposent un algorithme appelé *Social PageRank* dans un graphe social impliquant les utilisateurs en plus des documents et des annotations sociales [Bao et al., 2007]. En plus des graphes sociaux comprenant des documents et des utilisateurs, Kashyap et *al.*, proposent une personnalisation du PageRank sur un graphe social étendu, comprenant des utilisateurs, des groupes d'utilisateurs et des requêtes [Kashyap et al., 2012].

Dans d'autres travaux de personnalisation, un score de pertinence social est calculé pour les entités du graphe social ([Karweg et al., 2011], [Khodaei and Shahabi, 2012]). Karweg et *al.* introduisent deux facteurs sociaux pour estimer la pertinence sociale des entités du graphe [Karweg et al., 2011] :

– L'intensité d'engagement de l'utilisateur dans les activités sociales pour quantifier l'effort au moyen d'interactions sociales de l'utilisateur au sein du réseau, comme par exemple le clic sur un document retourné dans la liste des résultats, ou la recommandation d'un document ou d'un item à un ami ou un voisin social.

– Le score de confiance sociale, qui permet de déterminer l'intensité des relations de l'utilisateur avec les membres du réseau social, traduite comme étant une notion de confiance.

Le score social est donné par la combinaison de ces deux facteurs d'estimation de l'importance sociale des entités du graphe social.

Similairement, Khodaie et Shahabi proposent de mesurer l'importance sociale d'une entité (un document par exemple) au sein des réseaux sociaux par une combinaison multiplicative de trois facteurs calculés par des fonctions d'estimation de l'importance sociale $urf(u, u')$, $uaf(u', d)$ et $uwf(u')$ [Khodaei and Shahabi, 2012] :

– $urf(u, u')$ représente la fonction de mesure de pertinence sociale entre l'utilisateur initiateur de requête u et tout utilisateur u' relié au document d. Cette mesure est obtenue par le calcul d'une distance entre u et u'.

– $uaf(u', d)$ représente la fonction de mesure de l'activité de l'utilisateur u' en terme des actions de u' sur le document d (j'aime, partage, annotation, commentaire, etc.) ;

– $uwf(u')$ représente le poids social de l'utilisateur en terme de nombre d'amis, de suiveurs (*followers*). Il est calculé par les mesures classiques de degré de centralité (Centralité d'intermédiarité, de proximité, de prestige, de pouvoir, etc).

Ainsi, le score social $Score_S$ d'un document pour une requête est donné par la formule 3.1 :

$$Score_S(d, q) = \sum_{u'} urf(u, u') \times uaf(u', d) \times uwf(u') \tag{3.1}$$

3 Profil social

L'objectif dans cette section est de présenter les différentes approches de construction d'un profil social, du document ou de l'utilisateur, qui sont proposées dans les travaux de l'état de l'art.

3.1 Profil social du document

Le profil social de chaque document est représenté par un vecteur construit à partir des termes qui se trouvent dans les annotations sociales associées à ce document (cf. équation 3.2).

$$\overrightarrow{desc} : (t_1 : w_{desc,t_1}, ..., t_j : w_{desc,t_j}, ..t_n : w_{desc,t_n}) \tag{3.2}$$

où $w_{desc,t}$ est le poids social du terme t dans le profil du document d.

Ainsi, un poids social peut être calculé pour chaque terme du profil social du document en utilisant différentes pondérations [Noll and Meinel, 2007], [Xu et al., 2008], [Vallet et al., 2010], [Cai and Li, 2010], [Cai et al., 2010], [Xie et al., 2012].

Il peut être donné par une simple fréquence d'occurrence $tf_{desc,t}$ du terme dans les annotations sociales associées au document en question [Noll and Meinel, 2007].

Dans certains travaux d'état de l'art, les auteurs supposent que le profil social du document peut contenir un nombre important de termes. Ainsi, les auteurs proposent [Cai and Li, 2010], [Cai et al., 2010], [Xie et al., 2012] :

– de normaliser les $tf_{desc,t}$ par le nombre total des annotations associées au document et introduisent ainsi une nouvelle pondération appelée $tf_{desc,t}$ normalisé (*Normalized Term Frequency NTF*) [Cai and Li, 2010], [Xie et al., 2012] :

$$NTF_{desc,t} = \frac{|U_{d,t}|}{|U_d|} \tag{3.3}$$

où, $U_{d,t}$ représente l'ensemble des utilisateurs employant t pour annoter le document d et U_d représente l'ensemble des utilisateurs annotant d.

– d'introduire la normalisation de l'IDF [Cai et al., 2010] :

$$w_{desc,t} = tf_{desc,t} \times \log\left(\frac{|D|}{df_t}\right) \tag{3.4}$$

où, $tf_{desc,t}$ représente le nombre de fois où le document d a été annoté avec le terme t ; D est l'ensemble des documents du corpus et df_t est le nombre de documents annotés avec t.

D'autres pondérations sont utilisées : TF-IDF ([Xu et al., 2008], [Vallet et al., 2010]), à base du cosinus ([Xu et al., 2008], [Vallet et al., 2010]) et du modèle de pondération BM25 ([Bouadjenek et al., 2013], [Xu et al., 2008], [Vallet et al., 2010]).

Bouadjenek et *al.* proposent de générer une description sociale personnalisée d'un document vis-à-vis de l'utilisateur. Les auteurs se basent sur la construction d'une matrice de termes des annotations et d'utilisateurs pour représenter les termes employés par les utilisateurs sur les documents [Bouadjenek et al., 2013].

Le profil social d'un document peut être personnalisé selon chaque utilisateur l'annotant. Ainsi, pour chaque utilisateur formulant une requête, le profil social d'un document d selon l'utilisateur u (représenté par la formule 3.5) est donné sous la forme d'un vecteur de termes pondérés comme décrit dans la formule 3.6 :

$$\overrightarrow{desc_u} : (t_1 : w_{desc,u,t_1}, ..., t_j : w_{desc,u,t_j}, ..t_n : w_{desc,u,t_n}) \tag{3.5}$$

tel que, $w_{desc,u,t}$ est le poids du terme t dans le profil social du document d selon l'utilisateur u, il est calculé par une formule de pondération de type TF-IDF [Bouadjenek et al., 2013] :

$$w_{desc,u,t} = \frac{tf_{u,t}}{|T_{u,d}|} \times \log\left(\frac{|D_u| + 1}{|D_{u,t}|}\right) \tag{3.6}$$

avec :

– $tf_{u,t}$ est le nombre de fois où l'utilisateur u a employé le terme t pour annoter le document d,

– $T_{u,d}$ est l'ensemble des termes employés par u pour annoter d,

– D_u et $D_{u,t}$ représentent respectivement l'ensemble des documents annotés par u et l'ensemble des documents annotés par u avec le terme t.

Xu et *al.* et Vallet et *al.* pensent qu'il est plus intéressant d'inclure d'autres paramètres de pondération en plus des fréquences des termes pour évaluer leurs poids [Xu et al., 2008] [Vallet et al., 2010]. Ainsi, ils proposent d'appliquer différentes pondérations aux termes des annotations associées au document : une pondération à base d'un modèle de TF-IDF classique entre le profil social de l'utilisateur et celui du document, une pondération sur la base d'une fonction de similarité comme le cosinus ou une pondération sur la base d'un modèle en BM25.

3.2 Profil social de l'utilisateur

Il est parfois difficile d'identifier, de modéliser et d'exploiter les préférences et les centres d'intérêt des utilisateurs "*Most Web search engines in use fail to take advantage of the intentions, interests and preferences of their users*" [Pujol et al., 2003].

Avec l'arrivée des systèmes de *Folksonomies* et les différentes informations sociales issues de ces systèmes, plusieurs approches tentent de modéliser le profil de l'utilisateur par des vecteurs de centres d'intérêt basés sur les informations sociales de l'utilisateur [Al-Khalifa and Davis, 2006], [Michlmayr and Cayzer, 2007], [Au-Yeung et al., 2008], [Stoyanovich et al., 2008], [Szomszor et al., 2008], [Vallet et al., 2010], [Cai and Li, 2010].

Différentes informations peuvent être exploitées au sein d'un système de folksonomies pour construire le profil de l'utilisateur telles que les traces de l'utilisateur [Joachims, 2002a], les tweets [Yamaguchi et al., 2010], [Abel et al., 2011], les annotations et les relations sociales [Michlmayr and Cayzer, 2007], [Au-Yeung et al., 2008], etc.

Nous présentons essentiellement dans cette section les travaux exploitant les annotations et relations sociales pour la construction du profil social de l'utilisateur. L'ensemble des termes des annotations qu'un utilisateur a employés pour annoter des documents constitue une bonne représentation du profil social de l'utilisateur [Noll and Meinel, 2007], [Xu et al., 2008], [Vallet et al., 2010], [Cai and Li, 2010],

[Xie et al., 2012]. Différentes pondérations peuvent être appliquées à ces termes là. Le profil social est alors représenté par un vecteur de poids défini par :

$$\overrightarrow{Prof_u} : (t_1 : w_{u,t_1}, ..., t_j : w_{u,t_j}, ...t_n : w_{u,t_n}) \tag{3.7}$$

Le poids $w_{u,t}$ d'un terme t dans le profil social de l'utilisateur u peut être calculé par une simple fonction de pondération à base de fréquences d'occurrence de ce terme au niveau des annotations sociales de l'utilisateur ([Noll and Meinel, 2007]) :

$$w_{u,t} = tf_{u,t} \tag{3.8}$$

où t représente un terme dans les annotations sociales de l'utilisateur u et $tf_{u,t}$ est le nombre de fois où l'utilisateur u a employé t pour annoter des documents.

Dans d'autres approches, le poids social d'un terme du profil est normalisé par le nombre de documents annotés par l'utilisateur [Cai and Li, 2010], [Xie et al., 2012] :

$$w_{u,t} = tf_{u,t} \times \frac{1}{|D_u|} \tag{3.9}$$

où D_u représente les documents annotés par u.

Xu et *al.* [Xu et al., 2008] et Vallet et *al.* [Vallet et al., 2010] proposent ensuite différentes pondérations des termes du profil : une pondération de type TF-IDF (cf. équation 3.10) et une pondération basée sur le modèle BM25 [Robertson and Walker, 1994] (cf. équation 3.11) :

$$w_{u,t} = tf_{u,t} \times \log\left(\frac{|U|}{uf_t}\right) \tag{3.10}$$

où U représente l'ensemble des utilisateurs et uf_t est le nombre des utilisateurs employant t dans leurs annotations.

$$w_{u,t} = \frac{k_1 + 1) \times tf_{u,t}}{k_1 \times \left(1 - b + b \times \frac{ul}{agvul}\right) + tf_{u,t}} \times \left(\frac{|U| - uf_t + 0,5}{uf_t + 0,5}\right) \tag{3.11}$$

où ul est la taille du profil de l'utilisateur et $avgul$ est la taille moyenne d'un profil de l'utilisateur.

À la différence de [Xu et al., 2008], Vallet et *al.* choisissent un ensemble de termes dans les top k-documents dans la liste des documents annotés par l'utilisateur [Vallet et al., 2010].

Certaines approches proposent de sélectionner un sous-ensemble de termes pour représenter le profil social de l'utilisateur en exploitant la structure temporelle des données issues des actions d'annotation collaborative [Michlmayr and Cayzer, 2007]. Les auteurs montrent qu'il n'est pas suffisant de choisir les termes fréquents uniquement et proposent une approche qui se base sur la combinaison des termes qui co-occurrent dans les annotations pour construire un graphe social pondéré de co-occurrences de termes tel que, à chaque fois qu'un couple de termes (tags) qui co-occurre réapparaît, le poids du couple de terme augmente. Ainsi, le poids attribué au lien entre les termes co-occurrant augmente. Les top k-liens entre les termes ayant les plus grands poids sont ensuite sélectionnés pour représenter le profil de l'utilisateur.

Dans les travaux cités précédemment, les profils des utilisateurs ne tiennent pas compte du voisinage social de l'utilisateur. L'exploitation des relations sociales enrichit le profil de l'utilisateur, comme le montrent Stoyanovich et *al.*, où la pertinence prédite d'un document donné peut être améliorée en explorant les actions du voisinage (relations sociales) de l'utilisateur [Stoyanovich et al., 2008].

Le contexte informationnel social de u est alors représenté par les termes pondérés des annotations de l'utilisateur u (profil social de u) et de ceux de ses voisins sociaux.

Schenkel et *al.* proposent d'identifier les voisins sociaux de l'utilisateur par le biais d'une fonction de similarité des relations sociales de l'utilisateur (profil du voisinage social de u), tel que les voisins sociaux d'un utilisateur sont ceux annotant les mêmes documents que cet utilisateur et qui emploient les mêmes termes pour annoter les documents [Schenkel et al., 2008].

Pour identifier ces voisins sociaux de l'utilisateur annotant un document donné, les auteurs proposent de calculer une fonction de similarité basée sur les liens directs des utilisateurs connectés et les liens agrégés d'amitié entre les utilisateurs du réseau social qui n'ont pas de liens directs entre eux.

La fonction de similarité est ensuite utilisée comme une mesure de pondération des termes du contexte social de l'utilisateur incluant le voisinage social tel que, pour chaque document d on a :

$$\overrightarrow{Prof_{v_u,d}} : (t_1 : w_{d,u,t_1}, ..., t_j : w_{d,u,t_j}, ..., t_n : w_{d,u,t_n}) \tag{3.12}$$

où, $w_{d,u,t}$ est la combinaison des $tf_{d,u,t}$ du terme t employé par un utilisateur u annotant le document d et des $tf_{d,u',t}$ de chaque autre utilisateur u' du voisinage

social de u annotant d par t.

$w_{d,u,t}$ est calculé par la fonction de similarité des liens sociaux entre l'utilisateur u annotant un document d avec le terme t et les utilisateurs annotant d avec le même terme t.

$$w_{d,u,t} = \sum_{u' \in U} F_u(u') \times tf_{d,u',t} \tag{3.13}$$

avec :

– $tf_{d,u',t}$ est le nombre de fois où l'utilisateur u' a annoté le document d avec le terme t,

– $F_u(u')$ fonction de similarité des liens d'amitié de l'utilisateur :

$$F_u(u') = \alpha \times \frac{1}{|U|} + (1 - \alpha) \times sim(u, u') \tag{3.14}$$

– $sim(u, u')$ est la similarité entre deux utilisateurs u et u' en nombre de termes employés en commun pour annoter les mêmes documents :

$$sim(u, u') = max_{chemin \quad u=u_0,\ldots u_k=u'} \prod_{x=0}^{k-1} O(u_x, u_{x+1}) \tag{3.15}$$

– $O(u, u')$ est une fonction qui calcule la proportion entre les termes employés en commun par u et u' pour annoter des documents et l'ensemble de tous les termes des annotations de chaque utilisateur u et u'.

3.3 Discussion

Dans l'exploitations des annotations sociales pour la construction du profil et du contexte social de l'utilisateur, nous avons vu dans les travaux cités précédemment que le poids d'un terme au sein du profil et du contexte social peut être obtenu par différentes pondérations (pondération basée sur le TF simple, un TF normalisé (NTF), sur un modèle de type TF-IDF, sur le modèle BM25, etc.). Schenkel et *al.* supposent que les voisins sociaux sont les utilisateurs ayant des activités d'annotation similaires : les utilisateurs annotant le même document en employant fréquemment les mêmes termes. Les auteurs n'exploitent pas la notion d'utilisateurs reliés directement à l'utilisateur en question pour représenter le voisinage social de l'utilisateur [Schenkel et al., 2008].

Contrairement à Schenkel et *al.* nous pensons que les voisins sociaux d'un utilisa-

teur sont ceux qui sont reliés à l'utilisateur. Ainsi, il peut être intéressant de traiter le contenu textuel dans le contexte du voisinage social de l'utilisateur de la même manière que le contenu textuel d'un document, ainsi la pondération des termes du profil et du contexte du voisinage est appliquée au sein du modèle de RSPI proposé par [Schenkel et al., 2008].

4 Intégration des informations sociales

Une fois les informations sociales extraites et représentées, se pose la question de leur intégration dans le modèle de RI, particulièrement le niveau où elles peuvent être impactées. Dans cette section, nous présentons des travaux de l'état de l'art consacrés à l'intégration des informations sociales en distinguant ces différents niveaux.

4.1 Indexation sociale

L'intégration des informations sociales au niveau de l'indexation revient à considérer que les profils sociaux des documents, c'est-à-dire leurs indexations sont personnalisés pour chaque utilisateur.

Bouadjenek et *al.* proposent de calculer un score de similarité entre la requête et l'indexation sociale personnalisée du document pour un utilisateur donné [Bouadjenek et al., 2013] :

$$Score_S(desc_u, q) = sim_{Soc}(desc_u, q) \tag{3.16}$$

où, le poids d'un terme dans la description sociale du document pour un utilisateur initiateur de la requête est donné par la formule 3.6 (cf. section 3.1).

4.2 Discussion

Nous pensons que les annotations sociales associées au document qui peuvent être utilisées pour construire son profil social ne sont pas suffisantes pour représenter un document. Le contenu textuel d'un document est généralement la source d'information la plus complète et la plus intéressante, dans l'objectif d'améliorer la qualité de l'indexation des documents.

4.3 Reformulation et expansion de requête

Au fil des décennies, un grand nombre de méthodes de recherche d'information a été développé pour aider les utilisateurs à reformuler ou à affiner leurs requêtes. Ces méthodes peuvent être classées soit comme automatiques soit comme interactives [Ruthven, 2003].

Dans la plupart des approches d'expansion de requêtes utilisant les informations sociales issues des réseaux sociaux, il est question d'identifier les meilleurs termes candidats à l'expansion de la requête de l'utilisateur. Ces termes sont souvent sélectionnés à partir des annotations sociales, en particulier des termes du profil social de l'utilisateur.

La requête est alors étendue par une combinaison de termes initiaux de la requête et ces termes candidats [Schenkel et al., 2008], [Bouadjenek et al., 2011], [Xie et al., 2012]. Dans la plupart des cas cela peut être par une simple combinaison linéaire [Xie et al., 2012].

Dans Xie et al., la requête initiale formulée par un utilisateur u est alors représentée par un ensemble de termes pondérés, comme montré dans l'équation 3.17 :

$$q_u = (t_1^{q_u} : w_{q_u,t_1}, ..., t_j^{q_u} : w_{q_u,t_j} ...., t_m^{q_u} : w_{q_u,t_m}) \tag{3.17}$$

où, $t_j^{q_u}$ représente un terme t dans la requête q_u de l'utilisateur u et $w_{q_u,t}$ est le poids du terme t dans la requête q_u de l'utilisateur u tel que $w_{q_u,t} = 1$

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 3.18 :

$$q'_u = (t_1^{q'} : w'_{q_u,t_1}, ..., t_j^{q'} : w'_{q_u,t_j} ...., t_n^{q'} : w'_{q_u,t_n}) \tag{3.18}$$

où, le poids d'un terme dans la requête étendue par les termes des annotations sociales $w'_{q_u,t}$ est donné par une combinaison linéaire des termes initiaux de la requête de l'utilisateur et les termes du profil social de cet utilisateur (cf. équation 3.19) :

$$w'_{q_u,t} = \alpha w_{q_u,t} + (1 - \alpha) w_{u,t} \tag{3.19}$$

où, $w_{u,t}$ est le poids du terme t dans le profil de l'utilisateur u. La pondération utilisée pour calculer $w_{u,t}$ est une pondération à base de TF normalisés par la taille du profil social de l'utilisateur (cf. section 3.2, équation 3.9).

Dans d'autres cas, la sélection des termes candidats qui peuvent être ajoutés à la requête initiale se base sur une mesure de similarité sémantique entre les termes en question et ceux de la requête initiale [Schenkel et al., 2008], [Bouadjenek et al., 2011].

Schenkel et *al.* proposent de sélectionner parmi l'ensemble des termes du contexte social de l'utilisateur un sous-ensemble de termes qui peuvent être employés pour étendre la requête de l'utilisateur. Ainsi, en plus de la dimension sociale sur laquelle les auteurs se basent pour construire un contexte social de l'utilisateur (cf. section 3.2), 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 [Schenkel et al., 2008].

Pour chaque document d, la requête d'un utilisateur u est ensuite étendue par les termes des annotations de u employés pour le document d.

$$Score_{social}(t,t') = max_{t' \in T} sim_{tags}(t,t') \times ws_{d,u,t'}$$

(3.20)

– $Score_{sim}(t,t')$ est le score de similarité entre un terme de la requête et un terme candidat pour l'expansion de la requête,

– $sim_{tags}(t,t')$ est la mesure de similarité entre le terme de la requête t et le terme d'annotation t', basée sur le nombre de co-occurrence des termes,

– df_t est le nombre de documents annotés par le terme t.

$$ws_{d,u,t} = \frac{(k_1+1) \times |U| \times w_{d,u,t}}{k_1 + |U| \times w_{d,u,t}} \times IDF_t \tag{3.21}$$

– $ws_{d,u,t}$ est la version normalisée du poids du terme t dans le document d relative à l'utilisateur u posant la requête q,

– k_1 est le paramètre de normalisation de la saturation (cf. formule BM25 [Robertson and Walker, 1994]),

– $w_{d,u,t}$ est le poids social du terme t pour un document d (cf. section 3.2, équation 3.13).

Les termes avec le plus grand score social sont ensuite ajoutés aux termes initiaux de la requête.

Similairement, Bouadjenek et *al.* utilisent l'algorithme *Social Sim Rank* proposé par Bao et al. ([Bao et al., 2007]) pour calculer la similarité sémantique $Sim_{sem}(t,t')$

entre chaque terme t de la requête et les termes t' des annotations sociales de l'utilisateur qui représentent son profil social [Bouadjenek et al., 2011]. En plus de la dimension sémantique, les auteurs proposent d'adapter l'algorithme *Social Page Rank* de [Bao et al., 2007] pour calculer une similarité sociale $Sim_{sociale}(t', t_j)$ entre les termes du profil et ceux de la requête. Le score social final $Score_S(t, t')$ d'un terme candidat pour l'expansion de la requête de l'utilisateur est donné par une combinaison linéaire des deux similarités sémantique et sociale :

$$Score_S(t, t') = \alpha Sim_{sem}(t, t') + (1 - \alpha)\frac{\sum_{t_j \in Prof_u}^{m} Sim_{sociale}(t', t_j) \times w_{u,t_j}}{m} \quad (3.22)$$

avec

- m est le nombre total des termes du profil ($Prof_u$) de u
- $w_{u,t}$ est le poids du terme t dans le profil de u.

Les top k-termes avec les meilleurs scores sociaux sont ensuite ajoutés à la requête de l'utilisateur.

4.4 Discussion

Nous considérons que l'expansion des termes de la requête initiale de l'utilisateur par ceux de son profil permet de rajouter un contenu textuel considérable aux quelques termes initiaux de la requête (qui généralement ne dépassent pas 3 termes). De ce fait, une pondération à base de $tf_{q',t}$ normalisée seulement par la taille nous semble insuffisante et qu'il serait intéressant de prendre en compte les différentes caractéristiques liées aux distributions des termes dans la requête.

4.5 Reclassement des résultats

La prise en compte du contexte social au niveau des fonctions de correspondance et calcul de score a permis une amélioration significative des résultats de la recherche par un re-classement des documents retournés [Kirsch, 2005], [Noll and Meinel, 2007], [Xu et al., 2008], [Vallet et al., 2010], [Cai et al., 2010], [Cai and Li, 2010], [Gou et al., 2010], [Wen et al., 2012], [Kashyap et al., 2012], [Khodaei and Shahabi, 2012].

Noll et Meinel, proposent de calculer un score de similarité entre le profil de l'utilisateur et celui du document par une multiplication des TF des termes du profil

de l'utilisateur et ceux du profil du document pour améliorer ainsi le classement d'une liste de documents retournés par le moteur de recherche [Noll and Meinel, 2007].

Dans une première catégorie de combinaison de scores, les approches proposées se basent sur une combinaison du score thématique et d'un score social entre les profils sociaux (profil social du document et profil social de l'utilisateur). Le score social calculé dans Xu et *al.* et Vallet et *al.*, est calculé suivant différents modèles et fonctions de pondération [Xu et al., 2008], [Vallet et al., 2010].

$$Score(d, q, u) = \alpha RSV(d, q) + (1 - \alpha) \times Score_S(desc, u) \tag{3.23}$$

où :

1. RSV représente le score thématique classique obtenu avec un moteur de recherche classique.
2. $Score_S(desc, u)$, le score social entre le profil social du document et celui de l'utilisateur, ce score social est basé sur une fonction de pondération avec un TF-IDF classique, sur un calcul de similarité et sur une fonction de pondération BM25.

A la différence des travaux cités précédemment, nous pensons qu'il est essentiel de considérer le texte des contenus des documents dans la personnalisation sociale de la recherche d'information. Nous pensons qu'il serait important de comparer le profil social de l'utilisateur au contenu des documents, en plus du score thématique, dans l'objectif de retourner les documents qui correspondent le mieux aux attentes de l'utilisateur selon son profil.

Dans une autre catégorie de travaux, le score social est calculé entre la requête de l'utilisateur et les annotations sociales. Le score final est donné par la formule suivante [Wen et al., 2012], [Bouadjenek et al., 2013] :

$$Score - final(d, q, u) = \alpha RSV(d, q) + (1 - \alpha) Score_S(desc, q) \tag{3.24}$$

où le score social $Score_S(desc, q)$ peut être :

– un score social d'une requête par rapport aux annotations associées à un document, sans tenir compte du profil social de l'utilisateur [Wen et al., 2012]. Il

est calculé à travers une mesure de similarité entre la requête et les annotations sociales du document, basé sur la probabilité qu'un terme de la requête soit dans les annotations d'un document.

– un score social d'une requête par rapport aux annotations associées à un document par l'utilisateur initiateur de la requête (cf. équation 3.16) [Bouadjenek et al., 2013]. Contrairement à Wen et *al.*, Bouadjenek et *al.* proposent de personnaliser le profil social du document pour l'utilisateur initiateur de la requête [Wen et al., 2012].

Dans le cas où les informations sociales sont utilisées comme un indicateur d'importance social au lieu d'informations à propos des profils sociaux, les travaux d'état de l'art proposent des modèles de RSPI basés sur la combinaison des résultats d'une recherche classique (documents retournés par un moteur de recherche classique) et les résultats d'une recherche sociale tenant compte de l'importance sociale des différentes entités du graphe social ([Kashyap et al., 2012], [Khodaei and Shahabi, 2012]). La combinaison des résultats est donnée par la formule 3.25 :

$$Score\_final(d,q) = \alpha RSV(d,q) + (1-\alpha)Score_S(d,q) \tag{3.25}$$

où le score social $Score_S(d,q)$ représente :

– soit le score obtenu par une recherche sociale d'information au sein des entités du graphe social ([Kashyap et al., 2012]).

– soit le score de l'importance sociale du document, calculée à travers les annotations associées à ce document et du poids social de l'utilisateur annotant le document, calculé à partir de ses annotations et ses interactions avec les utilisateurs annotant le document, comme détaillé dans la section 2 et l'équation 3.1 ([Khodaei and Shahabi, 2012]).

5 Réseaux sociaux dans les systèmes de recommandation et de filtrage collaboratif

Les travaux cités dans les sections précédentes montrent des résultats intéressants à propos de l'exploitation des réseaux sociaux dans les systèmes de recherche d'information. Dans cette section, nous présentons des recherches connexes dans les domaines de recommandation et de filtrage collaboratif.

5.1 Réseaux sociaux et systèmes de recommandation

Les utilisateurs ont de plus en plus recours aux services de recommandation pour retrouver des items qui peuvent les intéresser : pour la recommandation d'experts [Davoodi et al., 2012], de musique [Konstas et al., 2009], de films ou de livres [Bonhard and Sasse, 2006], [Golbeck, 2006], etc.

L'incorporation des informations sociales a permis d'améliorer les systèmes de recommandation et d'affiner les résultats [Ma et al., 2011], [Tan et al., 2011]. Plusieurs travaux abordent le domaine de la recommandation personnalisée dans les réseaux sociaux et montrent des résultats prometteurs des systèmes de recommandation combinés aux réseaux sociaux [Carmel et al., 2010b], [Guy et al., 2010], [Li et al., 2013], [Breuss and Tsagkias, 2014].

Comme décrits par Guy et *al.*, les réseaux sociaux introduisent de nouveaux types d'information tels que les annotations, les évaluations, les commentaires et les relations sociales, etc. qui peuvent être exploités pour améliorer les recommandations et réciproquement, les techniques de recommandation contribuent au succès du Web social en permettant de mieux représenter l'utilisateur sur la base de ses préférences et de ses centres d'intérêts [Guy et al., 2010].

Différents systèmes de recommandation personnalisée destinés au grand public ont vu le jour avec l'arrivée des réseaux sociaux. À la tête de ces systèmes vient le moteur de recommandation personnalisée StumbleUpon [2] qui permet de suggérer des ressources web à l'utilisateur à partir de l'analyse de ses recherches, de ses évaluations antérieures, de celles de ses amis et de celles des utilisateurs avec qui il partage des intérêts similaires avec l'hypothèse qu'un utilisateur est beaucoup plus intéressé par les items (informations, produits, etc.) recommandés par les utilisateurs de son voisinage social (ses relations sociales) [Guy et al., 2009], [Guy et al., 2010].

Certaines approches reposent sur la personnalisation du *PageRank* [Brin and Page, 1998] pour tenir compte des interactions des utilisateurs au sein des réseaux sociaux [Li et al., 2013], [Carmel et al., 2010b], [Guy et al., 2010].

En personnalisant le *PageRank* et en l'adaptant au cas des réseaux sociaux, Li et *al* s'intéressent à l'exploitation des liens entre un sujet donné et une communauté et entre l'utilisateur et le sujet en question. Les auteurs proposent un framework appelé "FRec" qui permet de recommander pour un profil d'utilisateur donné à travers un sujet d'intérêt donné, les utilisateurs influents liés à ce sujet et les communautés

2. `http://www.stumbleupon.com`

interactives de sujets fédérateurs [Li et al., 2013].

Dans la même catégorie, Guy et *al* et Carmel et *al.* proposent une approche basée sur les systèmes sociaux d'agrégation qui permettent d'agréger des relations sociales à travers les utilisateurs, les items, et les termes. Le profil utilisateur est modélisé par une liste pondérée de termes et de personnes reliés à l'utilisateur. Les termes avec les plus grands poids sont considérés comme étant les termes les plus probables à employer par l'utilisateur u pour annoter un document d. Le même principe est appliqué pour prédire les utilisateurs qui auront tendance à annoter un document d par un terme t et pour prédire les documents susceptibles d'être annotés avec un terme t par l'utilisateur u [Guy et al., 2010], [Carmel et al., 2010b].

Breuss et Tsagkias proposent d'exploiter les cercles sociaux de l'utilisateur dans les réseaux sociaux (amis, *followers*, etc.) pour étudier les interactions de l'utilisateur au sein du réseau social et identifier les comportements sociaux similaires entre les utilisateurs pour recommander des contenus susceptibles d'intéresser l'utilisateur [Breuss and Tsagkias, 2014].

5.2 Réseaux sociaux et systèmes de filtrage collaboratifs

L'un des systèmes de filtrage collaboratif les plus connus est *ReferralWeb*[3] qui permet la fouille de n'importe quel réseau social sur le Web. *ReferralWeb* a pour objectif de permettre aux utilisateurs de trouver l'information dite "fiable" en provenance d'un expert, aussi qualifié de source fiable d'information, car l'information retournée est celle retrouvée au sein du voisinage social de l'utilisateur : ses amis ou les amis de ses amis, ses collègues, etc.

Comme pour les systèmes de RI et de recommandation, de nombreux travaux de l'état de l'art abordent les systèmes de filtrage collaboratif adaptés au cas des réseaux sociaux et des contenus générés par les utilisateurs (CGU) sur ses réseaux. Les approches proposées dans ces travaux montrent qu'il est intéressant de bénéficier de l'incorporation des informations sociales dans les systèmes de filtrage collaboratif pour améliorer les résultats retournés [Liu and Lee, 2010] [Ferrara and Tasso, 2011], [Ye et al., 2011], [Cheng et al., 2012], [Yuan et al., 2013].

Dans leurs travaux, Fengkun et Hong montrent une amélioration dans les performances d'un système de filtrage collaboratif en intégrant des informations sociales

3. `http://www.cs.rochester.edu/users/faculty/kautz/referralweb/`

issues des réseaux sociaux. Les auteurs proposent de collecter des informations sociales à propos des évaluations de préférences des utilisateurs et de leurs relations sociales pour identifier des voisinages sociaux en combinant des groupes d'amis et des plus proches voisins de chaque utilisateur. Ils évaluent ensuite leur système de filtrage collaboratif sur les groupes de voisinages sociaux [Liu and Lee, 2010].

Ferrara et Tasso supposent que les termes utilisés par les individus du voisinage social de l'utilisateur, dans les ressources partagées qui se retrouvent dans le sujet d'intérêt de cet utilisateur peuvent être considérés comme étant des termes pertinents pour le sujet d'intérêt de celui-ci [Ferrara and Tasso, 2011].

Dans la même catégorie, Ye et *al.* et Cheng et *al.*, exploitent les informations sociales dans les systèmes de filtrage collaboratif pour identifier des points d'intérêt pour l'utilisateur, prenant en compte l'influence géographique en plus de l'influence sociale, notamment celle des amis, sur l'utilisateur en question [Ye et al., 2011], [Cheng et al., 2012].

En plus des informations sur la localisation basée sur les réseaux sociaux pour la recommandation des points d'intérêt, Yuan et *al.* proposent d'introduire la notion de temps. Leur approche montre une amélioration importante dans les résultats retournés [Yuan et al., 2013].

6 Évaluation en recherche sociale personnalisée d'information

L'élément clé dans l'évaluation en recherche sociale personnalisée d'information est la collection de test dédiée. Nous avons vu dans le chapitre précédent les principales caractéristiques d'une collection de test dédiée à l'évaluation des modèles de RSPI (cf. chapitre 2), et la nécessité de disposer d'une telle collection pour l'évaluation des modèles de RI intégrant la dimension sociale. Les requêtes et les jugements de pertinence sont construits manuellement ou générés automatiquement. Dans cette section nous nous intéressons d'une part, aux principales compétitions en RSPI se basant sur la génération manuelle des requêtes et les jugements de pertinence et d'autre part aux travaux qui proposent des méthodes de construction automatique des requêtes et les jugements de pertinence.

6.1 Compétitions en recherche sociale personnalisée d'information

Avec l'arrivée de la nouvelle dimension sociale, on ne trouve pas encore de compétition dédiée comportant des requêtes et des jugements de pertinence centrée utilisateur. Nous citons dans cette section les deux principales compétitions utilisées pour l'évaluation des modèles de RI exploitant les informations sociales.

TREC Microblog : est une compétition de RSPI. La première collection de test de TREC Microblog, appelée *Tweets2011*, a été publiée dans le cadre de la compétition TREC. Elle se base sur les informations sociales du réseau social *Twitter*. TREC Microblog est dédiée à la RSI au sein des microblogs (ou tweets). *Tweets2011*[4] contient 16 millions de tweets et 49 topics [Ounis et al., 2011]. Publiée en 2013, la collection *Tweets2013* contient 240 millions de tweets de 200 millions d'utilisateurs et 70 topics [Jimmy and Miles, 2013]. Une des principales tâches de cette compétition est de permettre une recherche sur *Twitter* en temps réel.

INEX Social book search : est une compétition de RSPI. La première collection de test d'INEX Social book search a été publié en 2011. Cette collection contient 2,8 millions de livres collectés sur le site *Amazon.com*, 380 profils d'utilisateurs et 375 topics extraits du forum de réseau social d'amateurs de littérature *LibraryThing* [Koolen et al., 2012]. La seconde version de la collection, construite en 2013 contient 375 topics et 4572 jugements de pertinence [Koolen et al., 2013]. En 2014, le nombre de topics atteint 680 avec 8.918 jugements de pertinence. Une des principales tâches proposées par cette compétition est la recherche sociale de livres : l'utilisateur recherche un livre qui correspond à ses centres d'intérêt. L'objectif est de comparer la recherche traditionnelle de livres par le contenu avec la recherche de livre exploitant les contenus générés par l'utilisateur (CGU) tels que les commentaires, les annotations, les avis ou les notes que l'utilisateur attribue aux ressources sur le Web.

Dans ces quatres principales collections de test de RSPI, le contenu textuel complet des documents n'est pas disponible.

4. `http://trec.nist.gov/data/tweets/`

6.2 Éléments de la collection de test en recherche sociale personnalisée d'information

L'approche classique de construction de collection de test se base sur l'utilisation des jugements dits "d'experts" présente deux limitations (cf. section 1.7.1, chapitre 2) liées aux coûts et au temps passés pour la construction des collections ainsi qu'à l'incapacité des juges à estimer les attentes de l'utilisateur [Voorhees, 2005].

Ces limitations ont motivé la recherche d'alternatives pour la construction des collections de test, d'où l'apparition d'approches de génération automatique des éléments de la collection de test. Ces approches reposent sur le comportement de l'utilisateur observable (clics, reformulation de requêtes, temps de réponse [Kelly and Teevan 2003]) et/ou déduit à partir de ses activités sur les réseaux sociaux (avis, commentaires, annotations, liens sociaux [Schenkel et al., 2008], [Vallet et al., 2010]).

Nous présentons les principaux travaux d'état de l'art traitant de la construction automatique d'une collection de test de RSPI. Il s'agit d'une collection de test avec en plus des documents, requêtes et jugements de pertinence, l'identification d'utilisateurs avec des requêtes et jugements de pertinence centrée utilisateur.

6.2.1 Requêtes

Différents travaux proposent une génération automatique des requêtes de l'utilisateur en exploitant les annotations sociales au sein des réseaux sociaux [Vallet et al., 2010], [Xu et al., 2008], [Sanderson, 2008].

En général, les requêtes sur le Web ne contiennent pas plus de trois termes dans 85% des cas [Vallet et al., 2010].

Dans Xu et *al.* les auteurs considèrent que les termes des annotations d'un utilisateur peuvent être des mots clés des requête de l'utilisateur. Les auteurs proposent d'utiliser chaque terme d'une annotation comme une requête. Les requêtes sont donc composées chacune d'un seul terme [Xu et al., 2008].

Vallet et *al.* supposent qu'une requête utilisateur avec un seul terme est insuffisante pour représenter le besoin d'information de l'utilisateur [Vallet et al., 2010]. Les auteurs divisent les annotations de chaque utilisateur en deux parties : (1) une partie pour la construction des requêtes et (2) une partie pour la génération du profil de l'utilisateur. Les auteurs proposent de générer des requêtes à trois termes

qui sont extraits des annotations sociales de l'utilisateur. Pour chaque document les auteurs sélectionnent les trois termes des annotations des documents qui sont les plus populaires pour construire une requête. L'approche proposée dans [Vallet et al., 2010] montre une amélioration des résultats de la recherche avec des requêtes à trois termes par rapport à celles contenant un seul terme [Xu et al., 2008].

En plus de considérer les termes des annotations associés aux documents annotés par les utilisateurs, Schenkel et *al.* proposent de construire des requêtes sur la base d'une fréquence minimale (entre 1000 et 2000) d'apparition au sein des annotations sociales des couples de termes de la requête dans les documents annotés [Schenkel et al., 2008].

6.2.2 Jugements de pertinence

Certains travaux montrent que la construction des jugements de pertinence manuels est très coûteuse [Allan et al., 2007], [Carterette et al., 2008]. Aussi, deux principales sources d'informations sont exploitées pour construire automatiquement ces jugements de pertinence : les logs et les résultats de l'activité d'annotation (*bookmarking*).

Jugements de pertinence basés sur les logs :

Plusieurs travaux supposent que les informations obtenues à travers les logs ou les clics des utilisateurs reflètent les jugements de pertinence centrée utilisateur.

Noll et *al.*, [Noll and Meinel, 2007] proposent d'exploiter les informations sur les clics des utilisateurs sur les pages web retournées par un moteur de recherche, pour refléter la pertinence centrée utilisateur. Les auteurs proposent une liste personnalisée de documents retournés avec leur approche de calcul de score entre les profils des utilisateurs et ceux des documents. Les listes retournées pour 70% des requêtes utilisateurs sont différentes de celles retournées par un moteur de recherche sans approche de personnalisation. Ils construisent ensuite la pertinence centrée utilisateur pour chacun des utilisateurs en se basant sur l'ordre de leurs clics sur les documents retournés.

Jugements de pertinence basés sur les activités d'annotation :

Des travaux proposent d'exploiter les annotations sociales pour la génération automatique de la pertinence des documents, en supposant que les documents annotés par les termes de la requête et / ou ceux représentant le profil de l'utilisateur, peuvent être des documents pertinents pour la requête de l'utilisateur.

Dans le cas de requêtes comportant un seul terme, Xu et *al.* supposent qu'une ressource ou un document est pertinent pour un utilisateur donné s'il a été annoté par cet utilisateur avec le terme de la requête de l'utilisateur en question [Xu et al., 2008].

De la même manière, Vallet et *al.* proposent de construire les jugements de pertinence centrée utilisateur à partir des annotations de l'utilisateur [Vallet et al., 2010]. Les auteurs proposent une approche plus restrictive que celle de Xu et *al.*, [Xu et al., 2008] en considérant qu'un document est pertinent pour un utilisateur s'il est à la fois pertinent pour la requête de l'utilisateur et pour le profil original de l'utilisateur. Les documents annotés par les termes de la requête de l'utilisateur et qui correspondent au profil de celui-ci, sont considérés comme documents pertinents pour la requête de l'utilisateur.

En plus des annotations sociales de l'utilisateur, Schenkel et *al.* proposent d'inclure le voisinage social de l'utilisateur pour générer les jugements de pertinence centrée utilisateur [Schenkel et al., 2008]. Les documents annotés par l'utilisateur et son voisinage social sont considérés comme pertinents pour la requête de l'utilisateur.

7 Conclusion

Dans ce chapitre, nous avons pu voir de nombreuses contributions qui ont montré que l'incorporation des informations en provenance des réseaux sociaux apporte une amélioration des résultats des systèmes de recommandation, de filtrage et de recherche d'information. Nous avons abordé par la suite en détail, l'état de l'art consacré aux réseaux sociaux dans la recherche d'information.

Nous avons présenté les différentes contributions suivant deux grands axes : les modèles de RSPI et la collection de test de RSPI. Dans la partie sur les modèles de RSPI, nous avons recensé les travaux connexes d'exploitation des informations sociales pour la génération du profil social de l'utilisateur et du document et d'intégration du profil social de l'utilisateur dans le modèle de RSPI.

Dans la partie collection de test de RSPI dédiée, nous avons vu différents travaux qui proposent de générer des requêtes utilisateur et des jugements de pertinence centrée utilisateur, pour chaque couple (requête, utilisateur).

Chapitre 4

Modèles de recherche sociale personnalisée d'information

1 Introduction

Les premiers travaux en recherche d'information se sont basés essentiellement sur l'appariement entre les documents et la requête utilisateur, indépendamment de l'utilisateur initiateur de la requête. Avec l'émergence des réseaux sociaux et la disponibilité de différentes informations sociales, les travaux de RI ont commencé à s'intéresser d'avantage aux utilisateurs initiateurs des requêtes dans l'objectif de mieux identifier leurs besoins d'information et d'améliorer ainsi les résultats retournés, notamment en les personnalisant.

Les approches récentes de RSPI proposent d'exploiter des informations supplémentaires comme par exemple les annotations sociales, les avis, les notes des utilisateurs, etc.

Dans notre travail, nous nous focalisons sur l'utilisation des annotations sociales dans le processus de recherche d'information. Nous proposons un modèle de recherche sociale et personnalisée d'information qui tient compte du profil de l'utilisateur construit à partir de ses annotations sociales.

2 Motivations

2.1 Désambiguïsation de requête

Les requêtes envoyées à un système de recherche d'information sont souvent ambiguës. Il est souvent difficile, voire impossible d'interpréter de manière précise le besoin d'information représenté par une requête formulée par l'utilisateur sur un système de RI. De plus plusieurs utilisateurs peuvent formuler la même requête

sous la forme d'une liste de quelques mots clés mais en ayant chacun des besoins d'information différents en fonction de leurs centres d'intérêts.

En reprenant l'exemple de la section 3 du chapitre 1 rappelé dans le tableau 4.1,

	Smartphone	Android
Requête	1	1
Utilisateur Alice	40	3
Amis de Alice	55	12
Utilisateur Bob	3	40
Amis de Bob	15	54
Document d_1	85	7
Document d_2	7	85
Document d_3	85	85

Tableau 4.1 – Requête, profils utilisateurs et documents

si on suppose que les deux termes de la requête ont la même importance, un système de RI non personnalisée devrait estimer que d_1 est tout aussi pertinent que d_2 et cela quel que soit l'utilisateur considéré. Cependant, en fonction de son profil et de ses centres d'intérêt, le besoin d'information de l'utilisateur est susceptible de porter plus sur "*Smartphone*" ou sur "*Android*".

Ainsi, le profil d'Alice montre qu'elle est plus intéressée par les *Smartphones* que par *Android*. Ainsi, son besoin d'information est probablement centré autour du mot clé smartphone avec une ouverture sur *Android*. Le terme de la requête smartphone devrait donc être considéré de manière prioritaire pour Alice par rapport au terme de la requête *Android*.

Bob quant à lui, est principalement intéressé par les systèmes d'exploitation basés sur *Android*. Par conséquent, son besoin d'information est probablement centré autour d'*Android*, et donc le terme de la requête *android* devrait être considéré de manière prioritaire par rapport au terme de la requête *smartphone*.

Finalement, dans notre exemple, un système de RPI devrait considérer d_1 comme plus pertinent que d_2 pour l'utilisateur *Bob*, et l'inverse pour *Alice*.

2.2 Contexte social de l'utilisateur

Un contexte informationnel social de l'utilisateur peut être extrait de ses actions et activités au sein du réseau social. Il peut être composé de différents types

d'informations : annotations, commentaires, notes, relations sociales, etc (cf. figure 4.1).

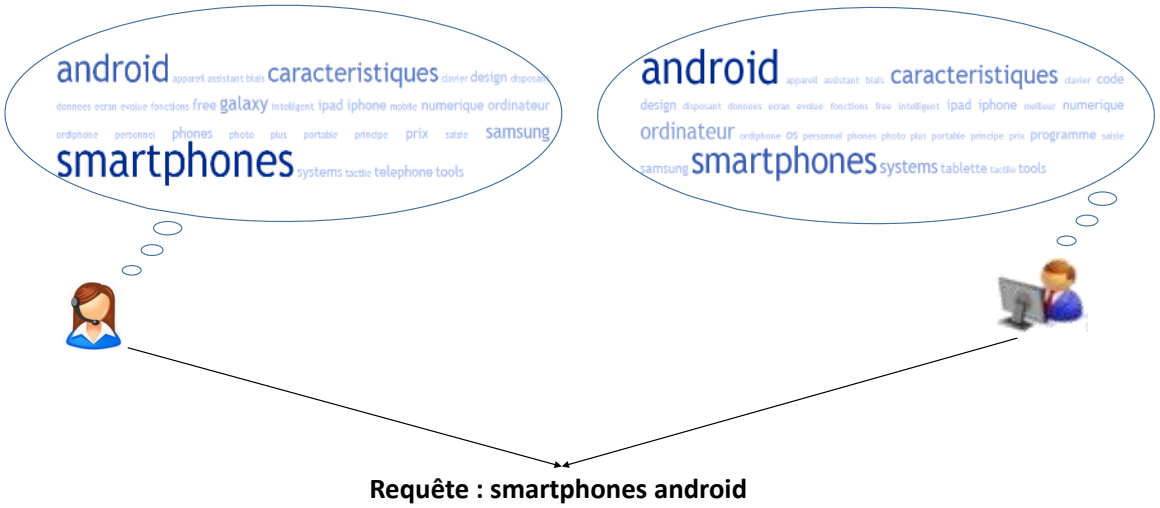

Figure 4.1 – Exemple : contexte social de l'utilisateur

Ce contexte informationnel social de l'utilisateur constitue une source d'information supplémentaire en plus des éléments classiques au sein d'un modèle de RI (requêtes et documents). Il est important de savoir à quel niveau du modèle de RSPI intégrer cette nouvelle source d'information et de quelle manière il faut l'intégrer.

3 Modélisation du contexte informationnel social de l'utilisateur (CIS)

Dans la suite, nous considérons que le contexte informationnel social de l'utilisateur est composé du profil social de l'utilisateur $PS(u)$, construit à partir de ses activités d'annotation et de celui de son voisinage social (ses relations sociales) $PV(u)$, construit à partir des annotations des utilisateurs de son voisinage.

3.1 Profil de l'utilisateur

Le profil social "$PS(u)$" de l'utilisateur u est représenté par les termes de ses annotations (cf. équation 4.1).

$$PS(u) = \{t \in T_z, a_z = < T_z, d, u > , a_z \in A(u)\} \tag{4.1}$$

où T_z est le sous-ensemble de termes apparaissant dans une annotation a_z et $A(u)$ est l'ensemble des annotations de l'utilisateur u.

Le profil social de l'utilisateur peut contenir plusieurs occurrences d'un même terme t. Ainsi, nous calculons une fréquence d'occurrence "$tf_{u,t}$" du terme t dans les annotations sociales de u qui représente le nombre de fois où l'utilisateur u a employé le terme t dans ses annotations pour annoter les documents.

Dans notre exemple (cf. Tableau 4.1), le profil social d'*Alice* est composé de 40 occurrences du terme ***Smartphone*** et de 3 occurrences du terme ***Android*** ($PS(Alice) = \{Smartphone, Android\}$; $tf_{Alice,Smartphone} = 40$; $tf_{Alice,Android} = 3$).

Alors que le profil de *Bob* est composé de 3 occurrences du terme ***Smartphone*** et 40 occurrences du terme ***Android*** ($PS(Bob) = \{Smartphone, Android\}$; $tf_{Bob,Smartphone} = 3$; $tf_{Bob,Android} = 40$).

3.2 Profil du voisinage de l'utilisateur

Des individus partageant les mêmes centres d'intérêt forment parfois un groupe et échangent des informations [Goh and Foo, 2008]. Ceci permet d'enrichir le profil social de l'utilisateur en l'étendant avec celui de son voisinage social. De même que pour le profil social de u, nous considérons que le profil du voisinage social "$PV(u)$" de l'utilisateur u peut être généré à partir des annotations des utilisateurs qui sont en relation sociale avec lui (cf. équation 4.2)

$$PV(u) = \cup_{u' \in U \ / \ (u,u') \in Rel} PS(u') \tag{4.2}$$

où Rel représente l'ensemble des liens sociaux entre les utilisateurs.

Le profil social de l'utilisateur peut contenir lui aussi plusieurs occurrences d'un même terme t. Ainsi, nous calculons une fréquence d'occurrence "$tf_{v_u,t}$" du terme t dans les annotations du voisinage de l'utilisateur, qui représente le nombre de fois où des utilisateurs du voisinage de u ont employé le terme t dans leurs annotations.

Dans notre exemple nous obtenons :

– Le profil du voisinage social d'*Alice* est composé de 55 occurrences du terme ***Smartphone*** et 12 occurrences du terme ***Android***,

– Le profil du voisinage social de *Bob* est composé de 12 occurrences du terme ***Smartphone*** et 55 occurrences du terme ***Android***.

4 Interprétations du contexte informationnel social pour la RI

Le contexte informationnel social contient des informations sur les centres d'intérêts de l'utilisateur. Dans l'objectif de l'exploiter pour la RI, nous distinguons deux interprétations possibles de ces informations que nous appelons "à propos" et "préférences".

– **Interprétation "à propos"** : les informations du contexte informationnel de l'utilisateur peuvent être interprétées comme un indicateur du contenu thématique des documents susceptibles d'intéresser l'utilisateur. Il a en effet déjà abordé ces thématiques dans ses activités d'annotation. On cherchera alors à retrouver les documents similaires au CIS, c'est à dire des documents dans lesquels la distribution des termes est proche de la distribution des termes dans le profil de l'utilisateur.

– **Interprétation "préférences"** : les informations du contexte informationnel social de l'utilisateur peuvent également être considérées comme un indicateur de préférences de l'utilisateur pour chaque terme. On cherchera alors à renforcer l'impact des termes "préférés" dans le calcul du score de pertinence des documents.

En reprenant l'exemple de l'utilisateur Alice, supposons que nous avons trois documents d_1, d_2, d_3 contenant chacun les termes Smartphone et Android comme rappellé dans le tableau 4.1

Avec **l'interprétation "à propos"**, on recherchera des documents similaires à la requête, comme ce qui se fait classiquement en RI, Alice préférera donc d_1 à d_2, même si les termes de la requête apparaissent globalement avec le même nombre d'occurrences dans les deux documents (85+7 = 7+85), car la proportion de t_1 et de t_2 dans son profil est plus proche de d_1 que de d_2. Alice préférera même d_1 à d_3 alors que le nombre d'occurrences total de t_1 et t_2 est plus important dans d_3 (85 + 85 = 170) que dans d_1 (85 + 7 = 92) car comparé à d_1, la distribution des termes dans d_3 est plus proche de celle dans le profil d'Alice.

Avec cette interprétation, un terme très peu important dans le profil sera considéré comme indésirable dans les documents recherchés : on pénalisera donc des documents contenant un nombre d'occurrences élevé de ce terme.

En calculant un score de similarité des documents de notre exemple par rapport à la requête posée par Alice, on obtiendra donc le classement suivant des documents :

$score_{RSI}(q, d_1) > score_{RSI}(q, d_3) > score_{RSI}(q, d_2)$ (cf. tableau 4.2).

Avec **l'interprétation préférences**, on cherchera à renforcer l'impact des termes importants dans le profil de l'utilisateur. Avec cette interprétation, Alice préférera d_1 (85+7) à d_2 (7+85) car le terme le plus important de son profil (*Smartphone*) sera prioritaire dans le calcul du score de pertinence des documents. Mais, contrairement à l'interprétation "à propos", Alice préférera d_3 (85+85) à d_1 (85+7), car le terme important du profil d'Alice (*Smartphone*) a le même nombre d'occurrences dans les deux documents, qui seront alors départagés par le nombre d'occurrences du terme *Android*, qui est plus fréquent dans d_3.

Si on considère que l'importance des termes dans le profil d'Alice traduit une préférence, alors le calcul de score retournera le classement suivant des documents :

$score_{RSI}(q, d_3) > score_{RSI}(q, d_1) > score_{RSI}(q, d_2)$ (cf. tableau 4.2).

On peut remarquer qu'un calcul de similarité entre les distributions des termes dans le document et dans le profil de l'utilisateur, n'est pas adapté à cette interprétation du profil. On préférera l'usage d'une fonction de type *multiplicatif*, qui augmentera l'impact des termes importants. On peut ensuite noter que dans le cas d'une interprétation "préférences", un terme qui apparaît dans le profil de l'utilisateur doit toujours être pris en considération même si son poids est faible. Contrairement au cas de la RI basée sur une similarité, où un terme est considéré comme indésirable si son poids dans la requête est faible.

Nous choisissons dans notre approche d'utiliser le contexte social de l'utilisateur comme indicateur de préférence de celui-ci.

5 Modèles de RSPI

Nous intégrons le contexte informationnel social de l'utilisateur, interprété comme indicateur de préférence, au sein du modèle de RSPI. Nous proposons de repondérer au sein du document ou de la requête les termes importants qui sont dans le CIS de l'utilisateur. La repondération des termes importants pour l'utilisateur dans

	t_1 Smartphone	t_2 Android	Similarité Rang : *à propos*	Multiplicative Rang : *préférences*
q	1	1		
$PS(Alice)$	40	3		
d_1	85	7	1^{er}	$2^{ième}$
d_2	7	85	$3^{ième}$	$3^{ième}$
d_3	85	85	$2^{ième}$	1^{er}

Tableau 4.2 – Interprétations préférences et à propos

le document et/ou dans la requête devrait améliorer les résultats retournés par le système pour la requête de l'utilisateur en renvoyant les documents pertinents pour chaque utilisateur selon son contexte informationnel social. Nous pensons qu'il est important de choisir un modèle de pondération tel que le BM25, qui est adapté au traitement de documents et de requêtes de tailles très variables, grâce à l'utilisation des versions normalisées des fréquences dans le TF, l'IDF et le QTF. Nous rappelons que le score d'un document pour une requête calculé dans le modèle BM25 est donné par la formule 4.3.

$$
\begin{aligned}
BM25(d,q) &= \sum_{t \in d \cap q} w_{d,t} \times w_{q,t} \\
&= \sum_{t \in d \cap q} \overbrace{\frac{(k_1+1)tf_{d,t}}{k_1(1-b+b\times\frac{dl}{avgdl})+tf_{d,t}}}^{TF_{d,t}} \times \overbrace{log\frac{N-df_t+0,5}{N-df_t}}^{IDF_t} \times \overbrace{\frac{(k_3+1)tf_{q,t}}{k_3+tf_{q,t}}}^{QTF_{q,t}}
\end{aligned}
$$

où $w_{d,t}$ est le poids du terme t dans le document d et $w_{q,t}$ est le poids de ce terme dans la requête q. Souvent, $w_{q,t}$ est égal à 1.

Nous proposons dans cette section deux approches d'intégration du contexte social de l'utilisateur au sein d'un modèle de RI, selon qu'il est combiné au document (cf. figure 4.2) ou à la requête (cf. figure 4.3).

Ceci nous a amené à définir plusieurs modèles de RSPI décrits dans les sections suivantes listés dans le tableau 4.3.

Dans ces modèles, nous intégrons le contexte social de l'utilisateur comme indicateur de ses préférences. En tenant compte de trois principaux niveaux de saturation

Niveau d'intégration du CIS dans une interprétation de type préférences	Modèles de RSPI
Intégration du contexte à l'indexation	Modèle $BM25F_S(d,q,u)$
Intégration du contexte à l'interrogation	Modèle $BM25S_{bin}(d,u)$ Modèle $BM25S_{tf}(d,u)$ Modèle $BM25S_{w}(d,u)$
	Modèle $BM25S_{FreqComb-bin}(d,q,u)$ Modèle $BM25S_{FreqComb-tf}(d,q,u)$ Modèle $BM25S_{FreqComb-w}(d,q,u)$
	Modèle $BM25S_{ScoreComb-bin}(d,q,u)$ Modèle $BM25S_{ScoreComb-tf}(d,q,u)$ Modèle $BM25S_{ScoreComb-w}(d,q,u)$

Tableau 4.3 – Tableau récapitulatif des différents modèles de RSPI proposés

des termes (saturation nulle, maximale et optimisée ou équilibrée) au sein du profil de l'utilisateur et au sein de la requête (par le biais du paramètre de contrôle de saturation k_3), nous proposons trois variantes de chaque modèle de RSPI intégrant le profil de l'utilisateur à l'interrogation (cf. figure 4.3).

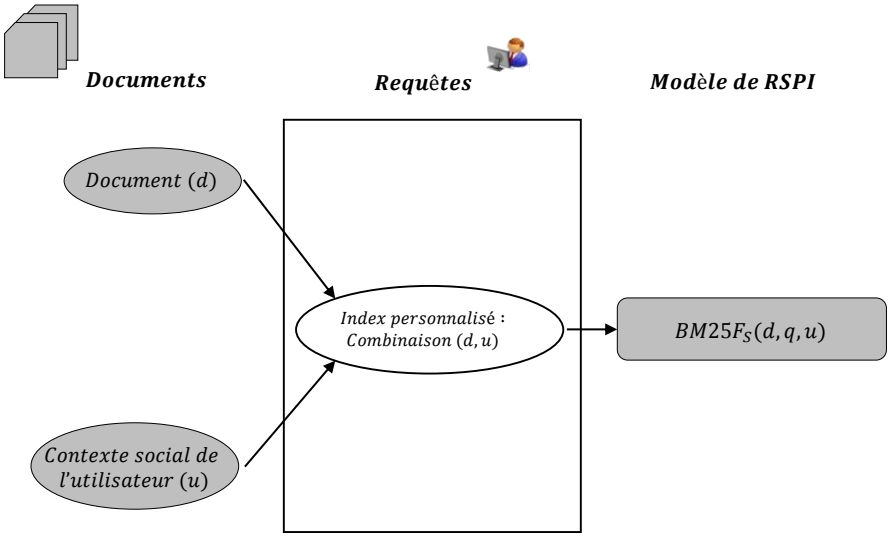

Figure 4.2 – Combinaison du contexte social de l'utilisateur au niveau des documents

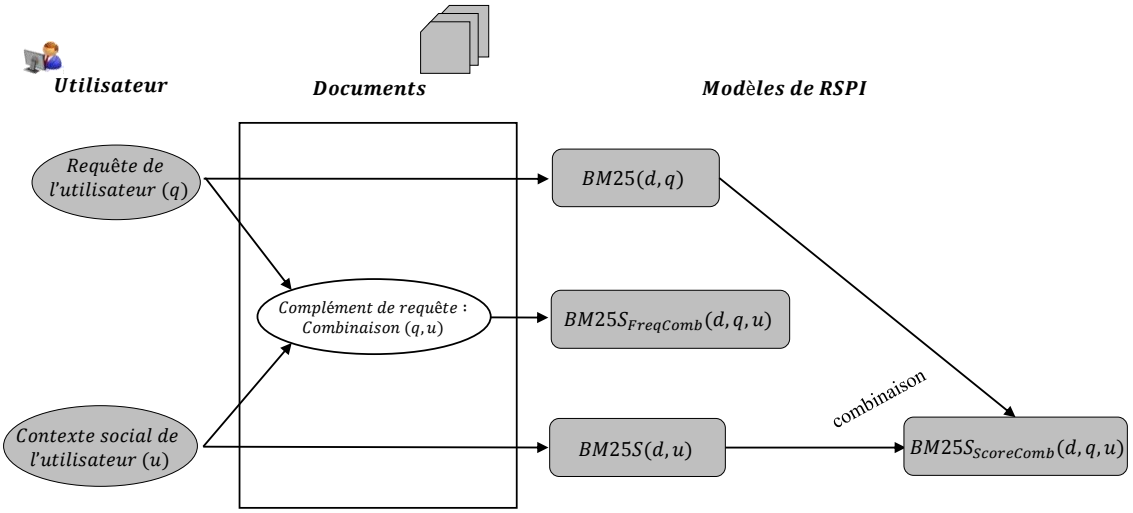

Figure 4.3 – Combinaison du contexte social de l'utilisateur au niveau de la requête

5.1 Personnalisation de l'indexation

Les annotations sociales peuvent enrichir la description des documents auxquels elles ont été associées par l'utilisateur. Dans notre approche de personnalisation de l'indexation, nous proposons d'utiliser le contexte social de l'utilisateur modélisé (cf. section 3) pour extraire depuis le document une description sociale personnalisée de ce document selon les termes du CIS de l'utilisateur. Pour cela, nous proposons d'identifier au sein du document les termes importants à repondérer et de combiner ensuite ces termes repondérés aux termes du contenu initial du document.

5.1.1 Repondération des termes du CIS

Nous avons proposé de repondérer les termes du document quand ils se retrouvent au sein du profil de l'utilisateur ([Bouhini et al., 2013a], [Bouhini et al., 2013b]). La repondération de ces termes permet d'accorder plus d'importance à ces termes qui représentent des informations à propos des centres d'intérêt de l'utilisateur.

5.1.2 Modèle de RSPI : $BM25F_S$

L'intégration du contexte informationnel social de l'utilisateur au niveau des documents (cf. figure 4.2) nous a conduit à proposer un premier modèle de recherche sociale personnalisée d'information, que nous appelons $BM25F_S$, basé sur le modèle de pondération $BM25F$ (cf. section 1.5, chapitre 2).

Dans ce modèle $BM25F_S$, nous proposons de générer un index personnalisé de documents par utilisateur, en deux étapes détaillées ci-après.

Description sociale personnalisée du document à partir du CIS de l'utilisateur :

Nous proposons de générer une description personnalisée du document par rapport au CIS de l'utilisateur et de combiner ensuite cette description personnalisée au contenu textuel du document. Cette étape est divisée en deux parties :

– La première consiste à identifier les termes important à re-pondérer. Il s'agit d'extraire depuis le document, les termes qui se trouvent aussi dans le profil de l'utilisateur. Ces termes sont considérés importants pour l'utilisateur et nécessitent d'être re-pondérés selon l'importance que l'utilisateur leur a accordé dans son profil.

– La deuxième est une re-pondération des termes du document qui constituent sa description sociale personnalisée. Ces termes obtiendront leur poids depuis le CIS de chaque utilisateur de sorte que quand un terme fait partie de la description sociale du document, le poids qu'on lui attribue sera celui du poids que ce terme obtient au sein du contexte social de l'utilisateur.

Combinaison de la description personnalisée du document avec son contenu :

Nous proposons de combiner le contenu du document avec la description sociale personnalisée du document par rapport au CIS de l'utilisateur, considérés comme étant trois différents champs d'information (profil social, profil du voisinage social et contenu initial du document). Nous générons, par conséquent, un index personnalisé des documents basé sur ces trois champs :

– *Le champ contenu du document d*, représenté par un vecteur de fréquences de termes dans le champ contenu du document (*field term frequencies* $ftfs_{d,t}$), qui est égale à la fréquence d'occurrences classique $tf_{d,t}$ du terme dans le document d :

$$ftfs_{d,t} = tf_{d,t} \tag{4.3}$$

– *Le champ description sociale personnalisée du document par rapport au profil de l'utilisateur* $PS(u)$, représenté par un vecteur de fréquences de termes du profil de l'utilisateur qui sont dans le document ($ftfs_{u,d,t}$).
Uniquement les poids des termes apparaissant à la fois dans le contenu du document et dans le profil social de l'utilisateur doivent être pris en considération :

$$ftfs_{u,d,t} = \begin{cases} tf_{u,t} & \text{si } tf_{d,t} > 0 \\ 0 & \text{sinon.} \end{cases} \tag{4.4}$$

où : $tf_{u,t}$ est la fréquence d'occurrence du terme t dans le profil social $PS(u)$ de u.

– *Le champ description sociale personnalisée du document par rapport au profil du voisinage de l'utilisateur* $PV(u)$, représenté par un vecteur de fréquences de termes du profil du voisinage de l'utilisateur et qui sont dans le document $ftfs_{v,d,t}$.
Similairement :

$$ftfs_{v_u,d,t} = \begin{cases} tf_{v_u,t} & \text{si } tf_{d,t} > 0 \\ 0 & \text{sinon.} \end{cases} \tag{4.5}$$

où : $tf_{v_u,t}$ est la fréquence d'occurrence du terme t dans le profil $PV(u)$ du voisinage social de u.

Chaque document est ensuite indexé par un vecteur de poids $ws_{u,d,t}$, avec $ws_{u,d,t}$ désignant une version personnalisée du poids $w_{d,t}$ du terme t dans le document d pour l'utilisateur u, calculée suivant la formule $BM25F$.

La fonction de pondération BM25F a été proposée par [Robertson et al., 2004] dans l'objectif d'utiliser une pondération basée sur le BM25 pour indexer des documents structurés composés de plusieurs champs (*fields*) : *titre*, *résumé*, etc. $BM25F$ est plus approprié que le $BM25$ pour l'indexation d'un document composé de différents champs.

Cette fonction de pondération a été étendue par Zaragoza et *al.*, dans l'objectif d'optimiser la normalisation de la taille de chaque champ du document [Zaragoza et al., 2004]. Nous avons choisi d'utiliser cette version.

Le score d'un document d pour une requête q calculé par la formule $BM25F$ dans [Zaragoza et al., 2004] est donné par la formule 2.16 que nous rappelons ci-dessous.

$$\overline{ftfs}_{d,t} = \frac{ftf_{d,t}}{1 + b_d \times (\frac{dl}{avgdl} - 1)} \tag{4.6}$$

$$\overline{ftfs}_{u,d,t} = \frac{ftfs_{u,d,t}}{1 + b_u \times (\frac{ul}{avgul} - 1)} \tag{4.7}$$

$$\overline{ftfs}_{v_u,d,t} = \frac{ftfs_{v_u,d,t}}{1 + b_v \times (\frac{vl}{avgvl} - 1)} \tag{4.8}$$

où :

– b_d, b_u et b_v sont des paramètres respectifs similaires au paramètre b (de BM25), pour chacun des champs que nous considérons ici respectivement, le champ contenu du document d, le champ description personnalisée du document par rapport au profil de l'utilisateur et le champ description personnalisée du document par rapport au profil du voisinage social de l'utilisateur.

– dl, ul et vl représentent ici les tailles des trois champs du document.

– $avgdl$, $avgul$ et $avgvl$ correspondent à la taille moyenne, à travers le corpus, des trois champs considérés.

Nous calculons un score $BM25F_S(d, q, u)$ d'un document d pour la requête q posée par l'utilisateur u.

$$\begin{aligned} BM25F_S(d,q,u) &= \sum_{t \in d \cap q} ws_{u,d,t} \\ &= \sum_{t \in d \cap q} TF_{S_{d,u,t}} \times IDF_t \times QTF_{q,t} \end{aligned}$$

où :

$$TF_{S_{d,u,t}} = \frac{ctf_{u,d,t}}{k_1 + ctf_{u,d,t}} \tag{4.9}$$

– IDF_t et $QTF_{q,t}$ représentent respectivement IDF_t et $QTF_{q,t}$ classiques dans la formule BM25 donné dans l'équation 4.3,

– $ws_{u,d,t}$ est le poids social du terme t dans le document d pour la requête q et l'utilisateur u,

– $ctf_{u,d,t}$, est la fréquence d'occurrences combinée des trois champs du document :

$$ctf_{u,d,t} = w_d \times \overline{ftfs_{d,t}} + w_u \times \overline{ftfs_{u,d,t}} + w_v \times \overline{ftfs_{v_u,d,t}} \tag{4.10}$$

– w_d, w_u et w_v sont des paramètres utilisés pour optimiser l'importance de chacun des champs : contenu du document, description personnalisée du document par rapport au profil de l'utilisateur et description personnalisée du document par rapport profil du voisinage de l'utilisateur.

5.2 Intégration du CIS aux documents : Positionnement et critiques

A notre connaissance, aucun des travaux antérieurs dans l'état de l'art ne propose d'intégrer le contexte informationnel social de l'utilisateur au niveau de l'indexation pour construire un index personnalisé des documents par utilisateur.

Dans l'approche la moins éloignée de la notre, Bouadjenek et *al* génèrent une description sociale personnalisée du document par rapport au profil de l'utilisateur, qui est construite à partir des termes des annotations que l'utilisateur a associé à ce document.

Au lieu d'une combinaison des fréquences d'occurrences des termes du contenu du document et de celui de sa description personnalisée comme proposé dans notre approche, les auteurs proposent une combinaison du score thématique de document pour une requête (RSV) et du score social de la description sociale personnalisée du document pour une requête). Comme montré dans Robertson et *al.*, nous considérons qu'une combinaison des champs d'information au niveau des fréquences d'occurrence est plus cohérente d'un point de vue théorique et expérimental qu'une combinaison de scores de chaque vecteur [Robertson et al., 2004]. Ce principe a été vérifié dans d'autres contextes, par exemple pour l'intégration du poids des balises pour la RI [Géry and Largeron, 2012], et des fréquences des mots traduits dans différentes langues [Li and Gaussier, 2012b].

Cette manière d'intégrer le contexte social de l'utilisateur au niveau des documents permet de personnaliser la recherche sociale d'information pour chaque utilisateur, en générant un index de documents centré utilisateur.

Cependant, l'approche de construction d'une représentation personnalisée des documents de l'index pour chaque utilisateur de manière distribuée peut être coûteuse ou au minimum délicate à mettre en œuvre dans le cas de très grands corpus de documents avec un grand nombre d'utilisateurs. C'est la raison pour laquelle nous avons proposé d'autres modèles où la personnalisation intervient au niveau de

la requête.

5.3 Personnalisation de requêtes

Dans un second modèle, nous proposons d'intégrer le contexte social de l'utilisateur à l'étape d'interrogation, au niveau de la requête et de la fonction de correspondance.

5.3.1 Repondération des termes du CIS

Très souvent, les termes de la requête ont une pondération binaire : le terme apparaît dans la requête (fréquence d'occurrence du terme $tf = 1$) ou n'apparaît pas ($tf = 0$). Dans le cas où le profil social est utilisé comme une requête, celle-ci est considérée comme une requête composée d'un grand nombre de termes, au lieu de quelques termes (par exemple 3 termes, dans certaines collections de test construites). De plus, cette "requête" peut alors contenir plusieurs occurrences d'un même terme. Par conséquent, il semble intéressant d'aller plus loin qu'une simple pondération binaire, en considérant les fréquences des termes de la requête, comme le font les fonctions de pondération du côté du document.

Comme la saturation des occurrences des termes peut jouer un rôle important dans les fonctions de pondération, nous proposons dans la suite de tenir compte de cet effet de saturation et d'étudier son impact sur le choix de la pondération à travers différentes manières de considérer les fréquences de termes (tf).

5.3.2 Impact de la saturation au niveau de la requête

Dans le modèle de pondération BM25, le paramètre permettant de contrôler la saturation au niveau de la requête est le paramètre k_3. Généralement, on peut distinguer trois niveaux de saturation selon les valeurs du k_3 :

– $k_3 = 0$: nous appelons ce cas ***binaire*** ***(bin)***. Il permet de négliger la valeur des tf des termes dans la partie requête. On parle dans ce cas d'une saturation maximale des termes au sein de la requête.

$$QTF_{q,t} = \frac{(k_3 + 1) \times tf_{q,t}}{k_3 + tf_{q,t}}$$

$$= \frac{tf_{q,t}}{tf_{q,t}}.$$

– $k_3 = 1000$ [Jones et al., 2000] : nous appelons ce cas ***fréquentiel*** (tf). Fixer la valeur du k_3 à 1000 ou à une grande valeur revient quasiment à prendre en compte directement la valeur du nombre d'occurrence du terme t dans la requête q ($tf_{q,t}$). Dans ce cas la saturation est nulle.

$$QTF_{q,t} \approx tf_{q,t} \tag{4.11}$$

– k_3 normalisé (ou fixé à 8) [Robertson et al., 1996] : nous appelons ce cas ***normalisé*** (w) car l'ajustement des valeurs du k_3 pour trouver une valeur optimale de celui-ci ou le fait de le fixer à 8, permet d'ajuster l'effet de saturation des termes de la requête.

5.3.3 Combinaison des requêtes et du CIS

Afin de tenir compte de ces trois façons de contrôler la saturation de la requête via le paramètre k_3, nous proposons trois modèles de recherche sociale personnalisée d'information, notés respectivement $BM25S$, $BM25S_{FreqComb}$ et $BM25S_{ScoreComb}$, basés sur le modèle de pondération $BM25$, que nous détaillerons par la suite [Bouhini et al., 2014].

– $BM25S(d, u)$: dans ce modèle, nous considérons que le contexte social de l'utilisateur peut être utilisé pour remplacer complètement la requête de l'utilisateur. Ce modèle retourne une liste classée de documents pertinents pour un utilisateur u en considérant son contexte informationnel social (cf. figure 4.3),

– $BM25S_{FreqComb}(d, q, u)$: ce modèle retourne une liste classée de documents pertinents pour un utilisateur u en considérant sa requête combinée à son contexte informationnel social au niveau des fréquences d'occurrence des termes (cf. figure 4.3),

– $BM25S_{ScoreComb}(d, q, u)$: ce modèle retourne une liste classée de documents pertinents pour un utilisateur u en combinant la requête à son contexte informationnel social au niveau des scores (cf. figure 4.3).

Contrairement aux requêtes qui contiennent très peu de termes, le contexte informationnel social de l'utilisateur lui, peut contenir des dizaines voire même des

centaines de termes. Pour cette raison, nous nous interrogeons sur l'importance de la distribution / pondération des termes du CIS de l'utilisateur au niveau de la requête. Nous pensons qu'il est beaucoup plus intéressant d'étudier l'impact des différentes variantes de ces trois modèles de RSPI qui prennent en compte le poids d'un terme du CIS au niveau de la requête de trois différentes manières : binaire (bin), fréquentielle (tf) et normalisé (w).

Chaque modèle de RSPI ($BM25S$, $BM25S_{FreqComb}$ et $BM25S_{ScoreComb}$) possède trois variantes selon ces trois pondérations. Nous présentons en détail les neuf variantes dans les sections suivantes.

5.3.4 Modèle de RSPI : $BM25S$

Dans le modèle de RSPI $BM25S$, nous proposons de remplacer la requête de l'utilisateur par son contexte informationnel social. Ceci nous a conduit à remplacer $QTF_{q,t}$ par $UTF_{u,t}$ dans le calcul du score BM25 donné par la formule 4.3 pour obtenir le score social $BM25S$. Le score social $BM25S$ d'un document pour un contexte social de l'utilisateur est donc défini par la formule 4.12.

$$\begin{aligned} BM25S(d,u) &= \sum_{t \in d \cap CIS(u)} w_{d,t} \times w_{u,t} \\ &= \sum_{t \in d \cap u} TF_{d,t} \times IDF_t \times UTF_{u,t} \end{aligned}$$

où :

$$w_{d,t} = TF_{d,t} \times IDF_t \tag{4.12}$$

$$w_{u,t} = UTF_{u,t} \tag{4.13}$$

$$UTF_{u,t} = \frac{(k_3+1) \times [w_u \times tf_{u,t} + w_v \times tf_{v_u,t}]}{k_3 + [w_u \times tf_{u,t} + w_v \times tf_{v_u,t}]} \tag{4.14}$$

– $TF_{d,t}$ et IDF_t représentent respectivement $TF_{d,t}$ et IDF_t classiques dans la formule BM25 donnée dans l'équation 4.3,

– $w_{u,t}$ est le poids d'un terme t dans le contexte de l'utilisateur u,

– w_u et w_v représentent respectivement deux paramètres déterminés expérimentalement et correspondant au poids du profil de l'utilisateur et au poids du profil du voisinage de l'utilisateur.

Selon les trois principaux niveaux de saturation définis dans la section 5.3.1, le modèle de RSPI $BM25S$ se décline en trois variantes :

1. $BM25S_{bin}(d, u)$: pour un $k_3 = 0$,
2. $BM25S_{tf}(d, u)$: pour un $k_3 = 1000$,
3. $BM25S_{w}(d, u)$: pour un k_3 optimisé,

5.3.5 Modèle de RSPI : $BM25S_{FreqComb}$

Dans le modèle de RSPI $BM25S_{FreqComb}$, nous proposons de combiner au niveau des fréquences d'occurrence les termes de la requête de l'utilisateur avec les termes de son contexte informationnel social.

Nous proposons d'équilibrer l'importance des termes du profil de l'utilisateur par rapport aux termes de la requête grâce au paramètre w_u, représentant le poids du profil social de l'utilisateur que nous déterminons expérimentalement.

L'importance des termes du profil social de l'utilisateur par rapport à ceux du profil du voisinage social de l'utilisateur est équilibrée par un paramètre w_v qui représente le poids du profil du voisinage de l'utilisateur et qui est déterminé expérimentalement.

$$BM25S_{FreqComb-bin}(d, q, u) = \sum_{t \in d \cap q} TF_{d,t} \times IDF_t \times QTF_{S_{q,u,t}} \tag{4.15}$$

où :

$$QTF_{S_{q,u,t}} = \frac{(k_3 + 1) \times [tf_{q,t} + w_u \times tf_{u,t} + w_v \times tf_{v_u,t}]}{k_3 + [tf_{q,t} + w_u \times tf_{u,t} + w_v \times tf_{v_u,t}]} \tag{4.16}$$

De même que pour le modèle précédent ($BM25S$), nous proposons trois variantes du modèle de RSPI $BM25S_{FreqComb}$, en fonction des trois niveaux de saturation des termes de la requête.

$$BM25S_{FreqComb-bin}(d, q, u) \tag{4.17}$$

$$BM25S_{FreqComb-tf}(d, q, u) \tag{4.18}$$

$$BM25S_{FreqComb-w}(d,q,u) \tag{4.19}$$

5.3.6 Modèle de RSPI : $BM25S_{ScoreComb}$

Dans le modèle de RSPI $BM25S_{ScoreComb}$, nous proposons de combiner le score de correspondance d'un document pour la requête binaire de l'utilisateur et le score de correspondance du document pour le contexte informationnel social de l'utilisateur.

En considérant les trois variantes correspondants à différentes valeurs du paramètre k_3 réglant la saturation, nous introduisons trois variantes du modèle de RSPI $BM25S_{ScoreComb}$ définies dans les équations 4.20, 4.21 et 4.22. Nous proposons d'équilibrer l'importance du score social par rapport au score thématique de document pour une requête (RSV) grâce au poids w_u du profil de l'utilisateur que nous déterminons expérimentalement.

$$BM25S_{ScoreComb-bin}(d,q,u) = RSV(d,q) + w_u \times BM25S_{bin}(d,u) \tag{4.20}$$

$$BM25S_{ScoreComb-tf}(d,q,u) = RSV(d,q) + w_u \times BM25S_{tf}(d,u) \tag{4.21}$$

$$BM25S_{ScoreComb-w}(d,q,u) = RSV(d,q) + w_u \times BM25S_{w}(d,u) \tag{4.22}$$

5.4 Intégration du CIS aux requêtes : Positionnement et critiques

Dans notre approche de combinaison du contexte informationnel social au niveau des requêtes, nous avons proposé différentes manières d'exploiter le contexte social de l'utilisateur et la requête de l'utilisateur.

Nous pensons qu'il est intéressant de s'interroger sur l'impact du contexte social de l'utilisateur lorsque celui-ci est utilisé comme indicateur de préférence de l'utilisateur pour chaque terme dans son contexte social.

Dans une première proposition (correspondant au modèle de RSPI $BM25S$), nous utilisons le profil de l'utilisateur pour remplacer sa requête. Nous calculons donc un score social du document uniquement par rapport au profil de l'utilisateur : $BM25S(d,u)$. Dans certains travaux de l'état de l'art le score social du document

pour un profil de l'utilisateur est basé sur un score de correspondance entre la description sociale du document (annotations sociales associées au document) et le profil social de l'utilisateur. Ainsi, le contenu du document n'est pas pris en compte [Noll and Meinel, 2007], [Xu et al., 2008], [Vallet et al., 2010]. Noll et Meinel proposent de calculer un score à base de TF.IDF classique entre le profil du document et celui de l'utilisateur ($TF.IDF(desc, u)$) et ne prennent pas en compte la normalisation par la taille. Xu et *al.* et Vallet et *al.* adoptent différentes fonctions de correspondance et calcul de score basées sur BM25 ($BM25(desc, u)$) aussi qu'une fonction de correspondance reposant sur cosinus (cf. chapitre 3 sections 4.1 et 4.5).

A la différence de ces approches proposées dans les travaux antérieur de l'état de l'art, nous pensons qu'il est important de tenir compte du contenu du document pour un score social de correspondance entre le profil de l'utilisateur et le document. De plus, dans notre approche nous proposons d'étudier différents niveaux de saturation (*binaire*, *fréquentiel* et *optimisé*) au niveau de la requête remplacée par le profil de l'utilisateur.

Nous avons considéré ensuite qu'il serait intéressant de combiner les termes de la requête avec ceux du contexte informationnel social de l'utilisateur et ce, de différentes façons :

– par combinaison au niveau des fréquences d'occurrences, proposée dans le modèle de RSPI $BM25S_{FreqComb}$. Une approche approche similaire a été proposé par [Xie et al., 2012] et [Cai and Li, 2010]. Xie et *al.* proposent une simple pondération linéaire des termes du contexte social à base de tf normalisé par la taille du contexte (cf. chapitre 3, section 4.1). Nous pensons que dans le cas où le contexte social, contenant un nombre important de termes, est utilisé comme requête ou combiné à la requête, il vaut mieux d'utiliser une pondération comme BM25 qui dispose de versions normalisée des fréquences TF, IDF et QTF.

– par combinaison au niveau des scores, proposée dans le modèle de RSPI $BM25S_{ScoreComb}$ comme celui introduit dans [Xu et al., 2008], [Vallet et al., 2010]. Cependant Xu et *al.*, combinent le score social de correspondance des profils sociaux (du document et de l'utilisateur) à un score thématique (RSV).

Par ailleurs, dans les travaux cités précédemment ([Vallet et al., 2010], [Xu et al., 2008], [Xie et al., 2012]), les auteurs n'étudient pas l'impact de la saturation côté requête complétée par les termes du contexte informationnel social (dans la combinaison des fréquences des termes et la combinaison des scores). De même que

dans notre première proposition $BM25S$, nous trouvons intéressant de tenir compte des niveaux de saturation et nous proposons donc dans notre approche différentes variantes des modèles de RSPI combinant la requête et le contexte social de l'utilisateur ($BM25S_{FreqComb}$ et $BM25S_{ScoreComb}$), correspondants aux trois principaux niveaux de saturation des termes.

6 Conclusion

Dans ce chapitre, nous avons présenté en détail les principales contributions théoriques de ce travail, à savoir les modèles de recherche sociale personnalisée d'information. Nous avons commencé par étudier comment il était possible d'interpréter le contexte informationnel social de l'utilisateur pour la recherche d'information et nous avons pointé deux principales interprétations possibles. Suite à cette étude, nous avons choisi d'utiliser l'interprétation "préférences".

Nous avons ensuite proposé une intégration du contexte social de l'utilisateur à deux niveaux : côté documents ou côté requêtes. Dans cet objectif, nous avons proposé différents modèles de RSPI qui permettent d'intégrer le contexte social :

– Au niveau des documents ($BM25F_S$), en permettant de générer un index personnalisé de documents par utilisateur. Cette approche n'a à notre connaissance jamais été explorée dans les travaux antérieurs :

$$BM25F_S(d,q,u) = \sum_{t \in d \cap q} TF_{S_{d,u,t}} \times IDF_t \times QTF_{q,t}$$

avec :

$$TF_{S_{d,u,t}} = \sum_{t \in d \cap q} \frac{w_d \times \overline{ftfs_{d,t}} + w_u \times \overline{ftfs_{u,d,t}} + w_v \times \overline{ftfs_{v_u,d,t}}}{k_1 + (w_d \times \overline{ftfs_{d,t}} + w_u \times \overline{ftfs_{u,d,t}} + w_v \times \overline{ftfs_{v_u,d,t}})}$$

– Au niveau des requêtes ($BM25S$, $BM25S_{FreqComb}$ et $BM25S_{ScoreComb}$), où nous avons proposé différentes variantes des modèles de RSPI selon le type de la saturation des termes de la requête. Ceci doit nous permettre d'étudier l'impact de cet effet de saturation au sein des requêtes, une problématique qui n'a pas

encore été traitée en recherche sociale personnalisée d'information :

$$BM25S(d,u) = \sum_{t \in d \cap u} TF_{d,t} \times IDF_t \times UTF_{u,t}$$

avec :

$$UTF_{u,t} = \frac{(k_3+1)[w_u \times tf_{u,t} + w_v \times tf_{v_u,t}]}{k_3 + [w_u \times tf_{u,t} + w_v \times tf_{v_u,t}]}$$

$$BM25S_{FreqComb}(d,q,u) = \sum_{t \in d \cap q} TF_{d,t} \times IDF_t \times QTF_{S_{q,u,t}}$$

avec :

$$QTF_{S_{q,u,t}} = \frac{(k_3+1)[tf_{q,t} + w_u \times tf_{u,t} + w_v \times tf_{v_u,t}]}{k_3 + [tf_{q,t} + w_u \times tf_{u,t} + w_v \times tf_{v_u,t}]}$$

$$BM25S_{ScoreComb}(d,q,u) = RSV(d,q) + w_u \times BM25S(d,u)$$

Chapitre 5

Collection de test de recherche sociale personnalisée d'information

1 Introduction

L'évaluation des modèles de recherche sociale personnalisée d'information (RSPI) ne peut se faire à l'aide des collections de test de RI classique qui ne comportent ni requêtes ni jugements de pertinence centrée utilisateur. Elles nécessitent donc des collections de test dédiées. Les collections de test de RSPI disponibles lorsque nous avons commencé cette recherche ne disposent pas de jugements de pertinence par utilisateur et ne sont donc pas satisfaisantes pour une évaluation de modèles de RSPI (cf. chapitre 2).

Nous avons donc construit une première collection de test de RSPI dont nous présentons dans ce chapitre les principales étapes de création. Cette collection $DelRSI1$, a été construite à partir des données du réseau social d'annotations collaboratives "*Delicious*". Nous avons ensuite filtré cette collection pour construire une seconde collection réduite que nous appelons $FDelRSI1$. Nous présentons également une autre collection de test de RSPI $DelRSI2$, qui est une extension de $DelRSI1$, basée sur un corpus de documents plus volumineux.

1.1 Requêtes centrées utilisateur

Comme expliqué dans le chapitre 2, dans la majorité des collections de test de RI, une requête a une seule interprétation possible. Cependant, dans la pratique, ce n'est pas toujours le cas. Une même requête textuelle, soumise à un SRI, peut correspondre à des besoins d'information différents. L'interprétation de la requête

peut même varier au cours du temps pour un même utilisateur.

Dans l'exemple de la section 3 du chapitre 1, la requête q : " *Smartphone Android*" correspond à deux besoins d'informations différents de deux utilisateurs Alice et Bob. Nous considérons que la requête q est associée à l'utilisateur qui l'a formulée. Nous définissons ainsi des couples <requête, utilisateur> exprimant le besoin d'information d'un utilisateur donné. Dans notre exemple, nous avons donc deux couples <requête, utilisateur> : (q, Alice) et (q, Bob). Par la suite, nous appelons ces couples des requêtes centrées utilisateur.

1.2 Jugement de pertinence centrée utilisateur

Un système classique de RI fournit une seule liste de documents pertinents pour une même requête textuelle, quel que soit l'utilisateur formulant cette requête.

Nous considérons que les systèmes de recherche sociale personnalisée d'information (RSPI) doivent fournir une liste de documents adaptée pour chaque utilisateur, compte tenu de son besoin d'information dans son contexte social.

En reprenant l'exemple du chapitre 1, Alice et Bob posent la requête *Smartphone Android* mais leur besoin d'informations est différent. Alice veut acheter un Smartphone à base de système Android et Bob cherche la documentation sur le développement des applications dans les Smartphones avec des systèmes Android (cf. chapitre 1). Il devrait y avoir des jugements de pertinence pour Alice et d'autres pour Bob, afin d'évaluer les résultats retournés par le système pour la requête d'Alice et celle de Bob et cela, même si la requête est identique. Une collection de test de RSPI doit donc proposer des jugements de pertinence centrée utilisateur.

2 Source de données : Delicious

Del.icio.us est un réseau social qui permet l'annotation collaborative des documents sur le Web. Un utilisateur peut y référencer des documents en leur associant des annotations sociales, c'est-à-dire un sac de mots clés appelés *tags* (termes). Une annotation a_z d'un document d sur le Web est un triplet composé d'un utilisateur u, d'un document d et d'un ensemble $T_z = \{t_1, t_2, ...t_z\}$ de tags (mots-clés) associés.

Sur *Delicious*, les utilisateurs peuvent avoir des liens avec d'autres utilisateurs (amis, collègues, etc.), construisant ainsi leur voisinage social, comme montré dans

la figure 5.1.

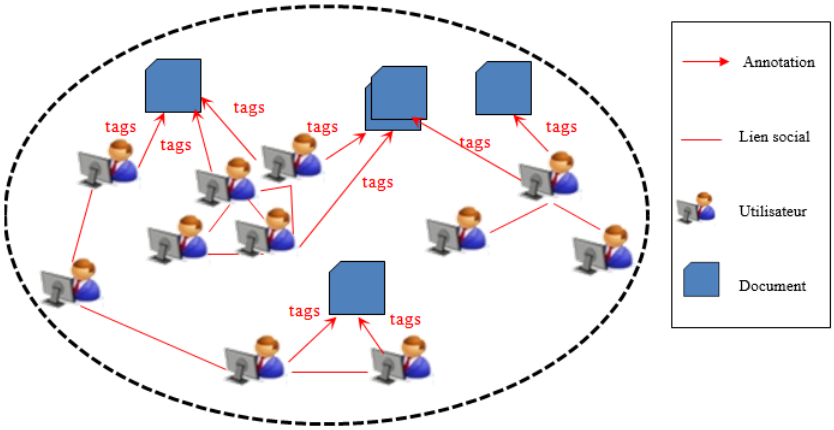

Figure 5.1 – Le réseau social *Delicious*

Delicious met à disposition deux principaux formats d'échange de données "***Flux RSS***" (Rich Site Summary) et "***Flux JSON***" (JavaScript Object Notation).

Nous avons choisi de collecter des données au format JSON pour construire notre collection de test.

La figure 5.2, présente un exemple de flux JSON d'une annotation sociale collectée sur *Delicious*. L'utilisateur *racum* a annoté le 4 octobre 2011 le document accessible à l'adresse *http ://fluidbaselinegrid.com/*, dont le titre est *"Fluid Baseline Grid-A sensible HTML5 and CSS3 development kit"*, avec les mots clés *responsive*, *design*, *web*, *css*, *css3*, *grid* et *fluid*.

[{"a": "racum", "d": "Fluid Baseline Grid - A sensible HTML5 and CSS3 development kit", "n": "", "u": "http://fluidbaselinegrid.com/", "t": ["responsive", "design", "web", "css", "css3", "grid", "fluid"], "dt": "2011-10-04T14:26:55Z"},..]

Figure 5.2 – Exemple d'une annotation au format JSON du réseau social *Delicious*

3 Construction de la collection de test de RSPI

La construction de notre collection de test en RSPI comporte trois étapes que nous détaillons dans cette section : la collecte de données publiques, la construction des requêtes et la construction des jugements de pertinence.

3.1 Collecte de données publiques

La collecte de données sur *Delicious* est initialisée par un ensemble de mots clés choisis manuellement et ciblant une thématique des documents annotés. *Delicious* propose plusieurs moyens d'accès à ces données concernant les annotations, les documents, les utilisateurs, etc., synthétisés dans le tableau 5.1. Pour la construction de notre collection de test de RSPI nous avons utilisé quelques unes de ces URLs, comme par exemple :

– l'URL de collecte des bookmarks par tags :
http ://feeds.delicious.com/v2/{format}/tag/tag[+tag+...+tag],
– l'URL de collecte des bookmarks par utilisateur :
http ://feeds.delicious.com/v2/{format}/{username},
– l'URL de collecte du voisinage social d'un utilisateur :
http ://feeds.delicious.com/v2/{format}/networkmembers/{username}.

3.2 Construction des requêtes des utilisateurs

Nous pensons qu'en formulant une requête, l'utilisateur a tendance à employer des termes qu'il a généralement aussi tendance à employer dans ses annotations sociales. Ainsi, en faisant l'hypothèse qu'un sous-ensemble de termes qui apparaissent conjointement fréquemment dans les annotations peut représenter les besoins d'information d'un utilisateur, nous proposons d'exploiter la fréquence d'occurrence des termes dans les annotations pour construire des requêtes des utilisateurs.

Nous avons choisi de limiter notre approche à la construction de requêtes de 2 ou 3 termes.

La première étape consiste à sélectionner les termes candidats pour constituer une requête. Nous supposons que les termes employés dans la même annotation sont sémantiquement liés. Nous utilisons le coefficient de Jaccard [Tan et al., 2005], permettant de mesurer le recouvrement de deux sous-ensembles, pour calculer la

Les URLs	***Flux JSON***
URL des bookmarks récents	- http ://feeds.delicious.com/v2/{format}/recent
URL des bookmarks par tags	- http ://feeds.delicious.com/v2/{format}/tag/tag[+tag+...+tag]
URL des bookmarks par utilisateur	- Signet d'un utilisateur spécifique : http ://feeds.delicious.com/v2/{format}/{username}
	- Signets privés d'un utilisateur spécifique : http ://feeds.delicious.com/v2/{format}/{username} ?private=key
	- Signets d'un utilisateur spécifique par tag : http ://feeds.delicious.com/v2/{format}/{username}/{tag}
	- Signets d'un utilisateur spécifique par tag : http ://feeds.delicious.com/v2/{format}/{username}/tag
	- Signets privés d'un utilisateur spécifique par tag : http ://feeds.delicious.com/v2/{format}/{username}/tag ?private=key
	- Signets privés d'un utilisateur spécifique par tag : http ://feeds.delicious.com/v2/{format}/{username}/tag ?private=key
URL pour les informations des utilisateurs	- Résumé information à propos de l'utilisateur : http ://feeds.delicious.com/v2/{format}/userinfo/{username}
	- Signets depuis l'inscription de l'utilisateur : http ://feeds.delicious.com/v2/{format}/subscriptions/{username}
URL des bookmarks pour le voisinage réseau	- Liste des membres du voisinage d'un utilisateur : http ://feeds.delicious.com/v2/{format}/networkmembers/{username}
	- Signets des membres du voisinage de l'utilisateur : http ://feeds.delicious.com/v2/{format}/network/{username}
	- Signets par tags des membre du voisinage de l'utilisateur http ://feeds.delicious.com/v2/{format}/network/{username}/tag
URL par ressources	- URL des bookmarks récents par URL d'une ressource http ://feeds.delicious.com/v2/{format}/url/{url md5}
Résumé d'informations à propos d'une url	- http ://feeds.delicious.com/v2/json/urlinfo/{url md5}

Tableau 5.1 – URL des bookmarks disponibles pour la collecte de données publiques sur *Delicious*

force de ce lien. Il s'agit de comparer la fréquence d'occurrence conjointe (nombre de co-occurrences) de couples de termes (pour les requêtes à deux termes) ou de triplets de termes (pour les requêtes à 3 termes), par rapport à la fréquence d'apparition globale (nombre d'occurrences) de chaque terme séparément.

Nous sélectionnons ensuite les "n" triplets de termes (ou couples de termes pour les requêtes à deux termes) ayant le coefficient de Jaccard le plus élevé (cf. figure 5.3).

Pour sélectionner les termes candidats, nous avons appliqué des filtres sur les données :

– ***Filtre des fréquences d'occurrence*** ***(***$Seuil_{TF}$***)*** : il est appliqué sur le nombre d'occurrence des termes au sein des annotations, que nous appelons $Seuil_{TF}$, tel qu'un terme t n'est utilisé dans la construction des requêtes que s'il a un nombre d'occurrence TF supérieur à $Seuil_{TF}$.
De cette manière nous éliminons les termes qui apparaissent rarement dans la collection de test.

– ***Filtres du seuil de Jaccard*** ***(***$Seuil_{2J}$ ***et*** $Seuil_{3J}$***)*** : dans ce type de filtres, nous fixons un seuil pour la valeur de Jaccard obtenue pour un couple de termes dans le cas des requêtes à deux termes ($Seuil_{2J}$), ou pour un triplet de termes dans le cas des requêtes à trois termes ($Seuil_{3J}$).
En appliquant le filtre de Jaccard, nous nous assurons que les termes choisis pour constituer la requête de l'utilisateur, co-occurrent suffisamment ensemble.

3.3 Collecte du contenu des documents sur le Web

Delicious offre la possibilité de collecter le contenu des documents annotés par les utilisateurs du réseau. Concrètement, il s'agit de pages web au format HTML. Nous collectons les contenus des documents du corpus qui correspondent aux annotations sociales collectées dans l'étape de construction de requêtes.

3.4 Construction des jugements de pertinence

Nous proposons une approche de construction des jugements de pertinence basée sur les annotations sociales des utilisateurs. En annotant des documents, les utilisateurs emploient des mots clés pour décrire de manière personnalisée les documents. Nous utilisons ces annotations pour définir la pertinence centrée utilisateur et la

pertinence globale. Pour la première, nous supposons qu'un document est pertinent pour une requête d'un utilisateur donné si ce document est annoté par au moins deux termes de la requête de cet utilisateur. Pour la seconde, nous considérons que la pertinence globale d'un document pour une requête (quel que soit l'utilisateur qui la pose) est dérivée de la pertinence centrée utilisateur puisque l'ensemble des documents pertinents pour cette requête est égal à l'union des documents pertinents de tous les utilisateurs ayant formulé cette requête (cf. figure 5.3).

– ***Filtre sur le nombre de documents pertinents (F_{DOC})*** : il est appliqué sur les documents pertinents par requête centrée utilisateur (couple <requête, utilisateur>) tel qu'on ne garde que les couples (requête, utilisateur) dont le nombre de documents pertinents est supérieur à un seuil F_{DOC}. Ce filtre réduira le nombre de couples (requête, utilisateur), par conséquence le nombre d'utilisateurs ayant des requêtes dans l'ensemble des requêtes centrées utilisateur sera réduit lui aussi.

– ***Filtre des valeurs par la précision moyenne (F_{AP})*** : notre évaluation se base sur le modèle de pondération BM25. Nous choisissons de filtrer les requêtes en fonction de la valeur de précision moyenne obtenue par un modèle de RI classique. Nous calculons la précision moyenne ($AP_{q,u}$) pour chaque couple <requête q, utilisateur u> avec le modèle BM25. Certains couples <requête q, utilisateur u> obtiennent des valeurs nulles ou très faibles avec la mesure d'évaluation AP, nous gardons uniquement les couples (q, u) avec une précision moyenne supérieure à un seuil F_{AP} fixé.

4 Formalisation de la proposition

4.1 Collecte des données de *Delicious*

Nous proposons de collecter les données des annotations sur le réseau *Delicious*, en procédant comme suit : Tout d'abord nous initialisons la collecte de données (bookmarks) en choisissant une liste de termes et nous collectons ensuite les annotations contenant ces termes. Une fois que nous avons collecté suffisamment de données correspondant à un nombre d'annotations à collecter fixé, nous récupérons l'ensemble des utilisateurs qui ont employé les annotations collectées précédemment puis celles des utilisateurs qui sont dans le voisinage social des utilisateurs ayant utilisé les premières annotations collectées.

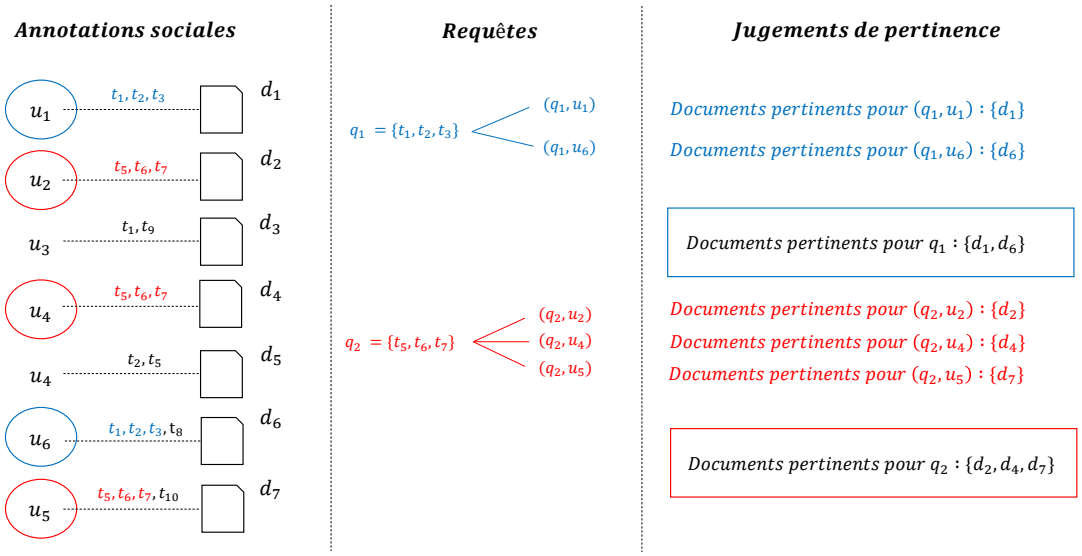

Figure 5.3 – Construction des requêtes et des jugements de pertinence

4.2 Construction de requêtes

Afin de retrouver les termes employés conjointement et fréquemment par les utilisateurs, nous proposons de construire une matrice de contingence de termes figurant dans les annotations employées par les utilisateurs.

On définit :

– $A(t) = \{a_z =< T_z, d, u >\in A/t \in T_z\}$: l'ensemble des annotations contenant le terme t.

– $|A(t) \cap A(t')|$: le nombre d'annotations contenant les termes t et t'.

Le choix des termes de la requête se fait sur la base des valeurs du coefficient de Jaccard pour chaque couple ou triplet de termes, par exemple on fixe le seuil de Jaccard $Seuil_{2J} = 0,5$ et le seuil de 100 requêtes à deux termes à construire.

En prenant le cas d'une requête q composée de deux termes ($q = \{t, t'\}$), la requête q est sélectionnée en fonction de la valeur du coefficient de similarité Jaccard :

$$jaccard_{t,t'} = \frac{|A(t) \cap A(t')|}{|A(t)| + |A(t')| - |A(t) \cap A(t')|} \tag{5.1}$$

Si $jaccard_{t,t'} > Seuil_{2J}$, alors la requête q est ajoutée à l'ensemble des requêtes

à deux termes $Q : Q = Q \cup q$, tel que :

Le même principe peut être étendu pour les requêtes à trois termes.

$$Jaccard_{t,t't''} = \frac{m_{tt't''}}{|A(t)| + |A(t')| + |A(t'')| - m_{tt'} - m_{t't''} - m_{tt''} + m_{tt't''}} \tag{5.2}$$

avec :

- $m_{tt't''} = |A(t) \cap A(t') \cap A(t'')|$
- $m_{tt'} = |A(t) \cap A(t')|$

4.3 Collecte de documents manquants pour les requêtes simulées

Pour collecter les données on utilise l'API de *Delicious*. Or l'API limite la collecte de données à un certain nombre d'annotations (ex : 100 annotations par terme, par utilisateur, par terme et utilisateur, par URL, etc.).

Dans l'objectif de collecter le maximum d'annotations concernées par les termes de la requête construite, nous collectons les documents annotés par les termes de la requête en collectant pour chaque requête, les bookmarks par combinaison des termes figurant dans cette requête.

4.4 Construction des jugements de pertinence

On considère que le corpus de documents est l'union des documents collectés qui sont annotés par les utilisateurs collectés dans la collection de test. Donc, d est pertinent pour le couple (q,u) si : $\exists a_z \in A, \quad \forall t \in q, t \in T_z, \quad a_z \quad = \quad < T_z, d, u >$.

Dans le cas des requêtes à trois termes, on considère que d est pertinent pour (q,u) s'il a été annoté par au moins deux termes sur trois de la requête.

5 Caractéristiques des collections de test de RSPI

5.1 Collection de test $DelRSI1$

La première collection de test que nous avons construite, appelée $DelRSI1$, contient 30224 documents, 70 utilisateurs, 79 requêtes globales avec 4685 jugements

de pertinence et 244 requêtes centrées utilisateur avec leurs 2730 jugements de pertinence centrée utilisateur correspondants (cf. tableau 5.2).

	Nombre	RI / RSPI
Documents	30.224	RI
Utilisateurs	70	RSPI
Requête Q	79	RI
Couples (requête, utilisateur) Q_{CU}	244	RSPI
Jugement de pertinence $Qrels_Q$	4.685	RI
Jugement de pertinence centrée utilisateur $Qrels_{CU}$	2.730	RSPI

Table 5.2 – Données finales de la collection de test $DelRSI1$.

Le tableau 5.2 indique également les informations figurant usuellement dans une collection de test de RI et celles qui sont spécifiques à une collection de RSPI. Les données de la collection de test $DelRIS1$, représentées dans le tableau 5.2 sont obtenues en suivant les étapes suivantes :

1. Collecte de données publiques : initialement nous avons collecté 29222 documents, 210 utilisateurs, 19815 termes et 69564 triplets.

2. Construction des requêtes et jugements de pertinence : les seuils appliqué dans la construction des requêtes et jugements de pertinence sont $Seuil_{TF} = 80$, $Seuil_{2J} = 0,5$ et $F_{DOC} = 10$.

   En appliquant ces seuils dans l'étape de construction des requêtes et des jugements de pertinence, nous avons obtenu 80 requêtes globales à deux termes avec 2922 jugements de pertinence. Les requêtes se déclinent en 2042 requêtes centrées utilisateur (couples <requête, utilisateur>) avec leurs 5554 jugements de pertinence centrée utilisateur.

3. Collecte des documents manquants : en complétant les jugements de pertinence par la collecte des documents manquants, nous avons obtenu au total 30984 documents dans la collection de test, 3983 jugements de pertinence globale et 6782 jugements de pertinence centrée utilisateur.

   Nous avons ensuite appliqué un seuil de 0,05% dans le filtre F_{AP} de sorte que nous gardons uniquement les couples <requête, utilisateur> obtenant une valeur de $AP > 0,05\%$ avec un modèle de RI BM25.

5.2 Collection de test $FDelRSI1$

Dans l'objectif d'éliminer les couples de requêtes trop spécifiques (ayant des valeurs de précision trop petites ou trop grandes) et les utilisateurs ayant très peu de requêtes, nous avons appliqué des filtres supplémentaires sur les données finales de la collection $DelRSI1$ présentées dans le tableau 5.2, de sorte que nous gardons uniquement les utilisateurs qui satisfont les conditions suivantes : avoir au minimum 5 requêtes par utilisateur et avoir une valeur de la moyenne des précisions moyennes MAP, obtenue avec le modèle de référence BM25, des requêtes de chaque utilisateur comprise entre $0,05\%$ et $\leq$0,7%.

Nous obtenons ainsi une liste de 10 utilisateurs avec 60 couples (requête, utilisateur) pour 12 requêtes globales restantes, comme indiqué dans le tableau 5.3 décrivant cette nouvelle collection $FDelRSI1$, obtenue par filtrage de la collection $DelRSI1$.

	Nombre	RI / RSPI
Documents	30.224	RI
Utilisateurs	10	RSPI
Requête Q	12	RI
Requête Q_{CU}	60	RSPI
Jugement de pertinence $Qrels_Q$	1095	RI
Jugement de pertinence $Qrels_{CU}$	764	RSPI

Table 5.3 – Données finales de la collection de test $FDelRSI1$.

5.3 Collection de test $DelRSI2$

La collection de test $DelRSI2$, construite en suivant la même procédure que pour $DelRSI1$, contient 161059 documents, 443 utilisateurs, 98 requêtes globales à trois termes avec 107923 jugements de pertinence et 1363 requêtes centrées utilisateur à trois termes avec leurs 31910 jugements de pertinence centrée utilisateur, comme indiqué dans le tableau 5.4.

Les données de la collection de test $DelRIS2$ sont obtenues en suivant les mêmes étapes que précédemment :

1. Collecte de données publiques : initialement nous avons collecté 141361 documents, 1720 utilisateurs, 60562 termes et 519962 triplets.

	Nombre	RI / RSPI
Documents	161.059	RI
Utilisateurs	443	RSPI
Requête Q	98	RI
Couples (requête, utilisateur) Q_{CU}	1.363	RSPI
Jugement de pertinence $Qrels_Q$	107.923	RI
Jugement de pertinence centrée utilisateur $Qrels_{CU}$	31.910	RSPI

Table 5.4 – Données finales de la collection de test $DelRSI2$.

2. Construction des requêtes et jugements de pertinence : les seuils appliqué dans la construction des requêtes et jugements de pertinence sont $Seuil_{TF} = 130$, $Seuil_{3J} = 0,015$ et $F_{DOC} = 10$.

   En appliquant ces seuils dans l'étape de construction des requêtes et jugements de pertinence, nous avons obtenu :

   – 100 requêtes globales à trois termes avec leurs 74626 jugements de pertinence,
   – 15010 requêtes centrées utilisateur à trois termes avec leurs 92519 jugements de pertinence centrée utilisateur.

3. Collecte des documents manquants : en complétant les jugements de pertinence pour les requêtes, nous avons obtenus 165578 documents dans la collection de test, 107980 jugements de pertinence globale pour les requêtes à trois termes et 137381 jugements de pertinence centrée utilisateur pour les requêtes à trois termes.

   Nous avons ensuite appliqué un seuil de 0,1% pour le filtre F_{AP} et nous obtenons les caractéristiques finales présentées dans le tableau 5.4

6 Évaluation

6.1 Résultats d'évaluation avec le modèle de référence

Dans un premier temps, nous évaluons le modèle BM25 dans un cadre de RI classique, en utilisant donc les requêtes globales des collections "$DelRSI1$" et "$DelRSI2$" avec leurs jugements de pertinence globale "$Qrels_Q$".

Les résultats d'évaluation sont présentés dans le tableau 5.5.

RI classique (requêtes et jugements de pertinence globale)	MAP	P[0.1]
Collection $DelRSI1$	0,1012	0,1775
Collection $DelRSI2$	0,0589	0,1603

Table 5.5 – Évaluation du modèle de référence $BM25$ avec les requêtes Q et les jugements de pertinence globale $Qrels_Q$ sur les collections $DelRSI1$ et $DelRSI2$.

Dans un second temps, nous évaluons le modèle classique BM25 en utilisant les requêtes centrées utilisateur et leurs jugements de pertinence centrée utilisateur correspondants.

Les résultats d'évaluation du modèle BM25 avec les couples <requête, utilisateur> Q_{CU} des collections $DelRSI1$ et $DelRSI2$ et leurs jugements de pertinence centrée utilisateur ($Qrels_{CU}$), sont présentés dans le tableau 5.6.

RSPI (requêtes et jugements de de pertinence centrée utilisateur)	MAP	P[0.1]
Collection $DelRSI1$	0,0308	0,0521
Collection $DelRSI2$	0,0108	0,0253

Table 5.6 – Évaluation du modèle de référence $BM25$ avec les requêtes Q_{CU} et jugements de pertinence sur les collections $DelRSI1$ et $DelRSI2$.

Les résultats obtenus par le modèle BM25 dans le cadre d'une évaluation en RI classique avec $DelRSI1$ sont du même ordre de grandeur que ceux habituellement constatés avec d'autres collections de test (MAP : 0,1012, P[0.1] : 0,1775), mais ils sont bien moins bons quand la taille de la collection augmente ($DelRSI2$, MAP : 0,0589, P[0.1] : 0,1603). Nous supposons qu'une collection de grande taille est une collection plus difficile à traiter par les systèmes de RI.

Par ailleurs nous constatons que les résultats sont très dégradés quand on se place dans le cadre plus difficile de la RSPI, en demandant au système de fournir des listes de documents adaptés à chaque utilisateur (requête centrée utilisateur et jugements de pertinence centrée utilisateur).

Le modèle BM25 de base n'étant pas capable de personnaliser les résultats utilisateur par utilisateur, il retourne en effet la même liste de documents pour une requête donnée à tous les utilisateurs. Cela explique les résultats très faibles obtenus :

- MAP : 0,0308 avec $DelRSI1$ et 0,0108 avec $DelRSI2$,
- $P[0.1]$: 0,0521 avec $DelRSI1$ et 0,0253 avec $DelRSI2$.

Cela montre aussi, la nécessité pour un modèle de RSPI de fournir des résultats personnalisés.

7 Conclusion

Dans ce chapitre nous avons détaillé les différents étapes de construction de nos collections de test $DelRSI1$, $FDelRSI1$ et $DelRSI2$, cette dernière a un nombre de documents cinq fois plus grand que dans $DelRSI1$. Nous avons ensuite présenté les résultats d'évaluation du modèle classique de RI (modèle BM25) sur nos collections de test de RSPI contenant des données centrées utilisateur (requêtes et jugements de pertinence centrés utilisateur).

Chapitre 6

Expérimentations

1 Introduction

Nous présentons dans ce chapitre les résultats d'évaluation des modèles de recherche sociale personnalisée d'information, proposés dans le chapitre 4, sur des collections de test de RSPI dédiées dont nous avons détaillé les caractéristiques dans le chapitre 5. Nous avons choisi d'utiliser le modèle BM25 comme modèle de référence.

2 Protocole expérimental

Nous avons évalué sur des collections de test de RSPI dédiées (cf. chapitre 5), nos modèles de RSPI intégrant le contexte social de l'utilisateur au niveau de l'indexation des documents ($BM25F_S$) et au niveau des fonctions de correspondance ($BM25S$, $BM25S_{FreqComb}$ et $BM25S_{ScoreComb}$). Les expérimentations ont été conduites sur le moteur de recherche Terrier version 3.5 [1], dans lequel nous avons activé l'outil de lemmatisation "Porter" [2].

Pour l'évaluation des résultats, nous avons utilisé le système "*TREC-Eval*" [3] (version 1.4, indépendamment de la plate-forme Terrier 3.5).

Nous avons employé les mesures d'évaluation suivantes (cf. chapitre 2, section 1.7.2) :

- MAP (*Mean Average Precision*) : la moyenne des précisions moyennes,
- $P[0.1]$: la précision à un taux de rappel de 10%.

1. `http://terrier.org/`
2. `http://terrier.org/docs/v3.5/javadoc/org/terrier/terms/PorterStemmer.html`
3. `http://trec.nist.gov/trec_eval/`

3 Personnalisation de l'indexation

Dans cette section, nous présentons tout d'abord les résultats d'évaluation du modèle de RSPI $BM25F_S$, intégrant le contexte informationnel social de l'utilisateur au niveau de l'indexation des documents, comparés aux résultats obtenus avec le modèle de référence BM25.

3.1 Collection de test utilisée

Nous avons utilisée la collection de test $FDelRSI1$, détaillée dans le tableau 5.3, qui est une version filtrée de la collection de test de RSPI "$DelRSI1$" (cf. tableau 5.2).

3.2 Résultats d'évaluation du modèle $BM25F_S$

3.2.1 Optimisation des paramètres

Le modèle $BM25F$ de Zaragoza et *al.* détaillé dans la section 1.5 du chapitre 4 sur lequel se base notre modèle de RSPI $BM25F_S$, est une version étendue du modèle $BM25F$ proposé par [Robertson et al., 2004]. Dans cette version étendue du $BM25F$, Zaragoza et *al.* supposent que les champs au sein d'un document structuré peuvent être de taille très variable. Ainsi, les auteurs proposent d'optimiser le paramètre b pour chaque champ (avec la meilleure valeur du paramètre k_1) au lieu d'optimiser un b global [Zaragoza et al., 2004].

Dans nos expérimentations, de la même manière que dans Zaragoza et *al.* nous avons optimisé pour chaque champ de $BM25F_S$ les valeurs des paramètres b et k_1 suivant une grille d'optimisation 2D. Dans un second temps, nous avons optimisé à nouveau un k_1 pour l'ensemble des champs en retenant les valeurs optimisées pour b pour chaque champ obtenues dans l'étape précédente, selon l'approche détaillée dans [Robertson et al., 2004].

1. Optimisation initiale du paramètre b et k_1 pour chaque champ :

   Dans l'optimisation du paramètre b, nous optimisons conjointement le couple des deux paramètres b et k_1 qui sont dépendants [Zaragoza et al., 2004] de la manière suivante : variation de $b \in [0, 1]$ avec un pas de 0,01 et $k_1 \in [0, 10]$ au sein de chaque optimisation du b par champ avec un pas de 1.

2. Optimisation finale de k_1 pour l'ensemble des champs :

   La valeur adaptée du paramètre k_1 est calculée indépendamment de chaque champ, en utilisant les valeurs optimales de b obtenues dans l'étape précédente, comme dans [Robertson et al., 2004]. La grille d'optimisation de k_1 est entre 0 et 10 ($k_1 \in [0, 10]$) avec un pas de 1.

3. Optimisation du poids de chaque champ :

   Une fois les valeurs de b pour chaque champ et de k_1 final obtenues, le poids w_f ($f \in \{d, u, v\}$) de chaque champ est ensuite optimisé de la manière suivante :
   - Le poids du champ "contenu" du document est fixé à 1 : $w_d = 1$,
   - Le poids de chacun des deux champs "profil social" et "profil du voisinage social" est optimisé de la manière suivante : $w_u \in]0, 1]$ avec un pas de 0,1 tout en mettant le poids w_v à 0. De la même manière, ensuite, $w_v \in]0, 1]$ avec un pas de 0,1 et $w_u = 0$.

Pour chaque utilisateur, nous optimisons les valeurs de b, k_1 et w_f. Le tableau 6.1 présente les moyennes et les écarts types des paramètres b et w_f optimisés sur l'ensemble des utilisateurs.

	Avec MAP	
	b : (moyenne, écart type)	w_f : (moyenne, écart type)
Champ "contenu" du document dans $BM25F_S$	(0,9 , 1)	(1 , 0)
Champ "profil utilisateur" dans $BM25F_S$	(0,34 , 0,20)	(0,44 , 0,18)
Champ "profil voisinage" dans $BM25F_S$	(0,19 , 0,12)	(0,22 , 0,13)

Table 6.1 – Paramètres optimisés du modèle $BM25F_S$.

Les moyennes et écarts types des valeurs de k_1 final re-optimisé sont respectivement 6,5 et 0,15. Ces valeurs des moyennes des paramètres optimisés (b, k_1) montrent que les valeurs obtenues pour les paramètres optimisés sont différentes de celles fixées par défaut dans Terrier (b = 0,75 et k= 1,2).

3.2.2 Résultats d'évaluation

Afin de montrer l'effet de l'intégration de différentes parties du contexte informationnel social de l'utilisateur au sein des documents, nous avons expérimenté deux

niveaux d'intégration.

– Niveau profil : correspond à l'intégration du profil de l'utilisateur dans le modèle $BM25F_S$, en gardant uniquement le poids du profil social de l'utilisateur w_u optimisé (cf. section 3.2.1) avec $w_v = 0$ et $w_d = 1$.

– Niveau profil et voisinage : correspond à l'intégration du profil de l'utilisateur et du profil de son voisinage social dans le modèle $BM25F_S$ en retenant les valeurs optimisées de w_v et w_u avec $w_d = 1$.

Le tableau 6.2 montre les résultats d'évaluation du modèle de référence et du modèle de RSPI $BM25F_S$ obtenus par les mesures de MAP globale (MAP) et de la précision globale de tous les couples <requête, utilisateur>.

$BM25$		$BM25F_S$			
		Profil_uniquement		*Profil_et_voisinage*	
MAP	$P[0.1]$	MAP	$P[0.1]$	MAP	$P[0.1]$
0,0226	0,0507	0,0282	0,0661	0,0289	0,0678

Table 6.2 – Résultats d'évaluation globale du modèle $BM25F_S$ sur la collection $FDelRSI1$.

Le tableau 6.2 montre des résultats d'évaluation du modèle de référence et du modèle de RSPI $BM25F_S$, obtenus en utilisant les configurations (*Profil_uniquement* et *Profil_et_voisinage*).

Les résultats présentés dans le tableau 6.2 montrent que le modèle de RSPI "$BM25F_S$" intégrant les données sociales de l'utilisateur ($CIS(u)$) permet d'améliorer les résultats de la RI.

En intégrant le profil de l'utilisateur uniquement, les résultats en MAP (0,0282) et en précision (0,0661) sont meilleurs que le modèle de référence. Ainsi, nous obtenons une amélioration relative de 14% en MAP par rapport au modèle de référence.

En ayant intégré, au sein du modèle de RSPI $BM25F_S$, le profil de l'utilisateur et celui de son voisinage, nous constatons que les résultats sont meilleurs que le modèle de référence et que le modèle $BM25F_S$ qui intègre seulement le profil de l'utilisateur uniquement : en MAP (0,0289) et précision (0,0678).

Le tableau 6.3 montre des résultats d'évaluation du modèle de référence et du modèle de RSPI $BM25F_S$, utilisateur par utilisateur.

En analysant les résultats d'évaluation utilisateur par utilisateur avec les mesures MAP et en précision $P[0.1]$ par utilisateur (cf. tableau 6.3), nous constatons

	$BM25$		$BM25F_S$			
			Profil_uniquement		*Profil_et_voisinage*	
	MAP	$P[0.1]$	MAP	$P[0.1]$	MAP	$P[0.1]$
u_1	0.0614	0.1310	0.0816	0.1426	**0.0819**	**0.1503**
u_2	0.0404	**0.2614**	0.0402	0.1265	**0.0416**	0.1203
u_3	0.0358	0.1076	**0.0486**	0.1438	0.0483	**0.1438**
u_4	0.0262	0.0922	**0.0278**	**0.0956**	0.0275	0.0948
u_5	**0.0287**	0.0569	0.0253	0.0484	0.0284	**0.0688**
u_6	0.0174	0.0529	**0.0199**	**0.0568**	0.0197	0.0550
u_7	**0.0183**	0.0207	0.0173	0.0298	0.0148	**0.0391**
u_8	0.0138	0.0296	0.0148	0.0316	**0.0156**	**0.0319**
u_9	0.0074	**0.0221**	**0.0093**	0.0215	0.0091	0.0188
u_{10}	0.0077	0.0095	0.0085	**0.0115**	**0.0103**	0.0101

Table 6.3 – Résultats d'évaluation utilisateur par utilisateur du modèle $BM25F_S$ sur la collection $FDelRSI1$.

qu'avec le modèle $BM25F_S$ qui intègre le profil de l'utilisateur seulement, sept utilisateurs sur dix obtiennent de meilleurs valeurs en MAP et en précision que le modèle de référence $BM25$.
En intégrant le profil et le voisinage de l'utilisateur au sein du modèle $BM25F_S$, nous obtenons une amélioration par rapport au modèle de référence $BM25$ et à la variante du modèle $BM25F_S$ qui intègre le profil de l'utilisateur seulement. Nous constatons aussi que pour huit utilisateurs sur dix une amélioration en MAP et en précision $P[0.1]$.

4 Personnalisation des requêtes

Nous présentons les résultats d'évaluation des modèles de RSPI intégrant le contexte de l'utilisateur au niveau de la fonction de correspondance par rapport au modèle de référence BM25. Dans l'objectif d'étudier les différentes combinaisons et manières d'intégrer du profil social au sein d'un modèle de RSPI, nous évaluons les trois modèles de RSPI suivants :

1. $BM25S(d, u)$, le modèle de RSPI intégrant le profil social de l'utilisateur comme étant une représentation d'un besoin d'information de l'utilisateur.
2. $BM25S_{FreqComb}(d, q, u)$ le modèle de RSPI intégrant le profil social de l'utilisateur comme étant un complément pour la requête initiale de l'utilisateur

avec une combinaison au niveau des fréquences d'occurrences des termes de la requête et du profil.

3. $BM25S_{ScoreComb}(d, q, u)$, le modèle de RSPI intégrant le profil social de l'utilisateur pour étendre la requête initiale de l'utilisateur par une combinaison du score classique du document pour la requête ($BM25(d, q)$) et du score social du document pour le profil de l'utilisateur ($BM25S(d, u)$).

4.1 Collection de test utilisée

Dans l'évaluation des modèles de RSPI intégrant le contexte informationnel social de l'utilisateur au niveau de la fonction de correspondance, nous utilisons la collection de test $DelRSI2$ (cf chapitre 5, tableau 5.4). Rappelons que la collection de test $DelRSI2$ comporte cinq fois plus de documents que la collection $DelRSI1$.

4.2 Résultats d'évaluation du modèle $BM25S$

Nous évaluons les trois variantes du modèle de RSPI "$BM25S$" ($BM25S_{bin}(d, u)$, $BM25S_{tf}(d, u)$ et $BM25S_{w}(d, u)$) par rapport au modèle de référence BM25.

4.2.1 Optimisation des paramètres

Nous optimisons les paramètres (b, k_1 et k_3) du modèle $BM25S$ et le poids du profil de l'utilisateur w_u de la manière suivante :

– dans un premier temps, nous optimisons le couple de paramètres dépendant b et k_1 en utilisant une grille d'optimisation 2D : $b \in [0, 1]$ avec un pas de 0,01 et $k_1 \in [0, 30]$ avec un pas de 1.
– en attribuant aux b et k_1 les valeurs optimales obtenues précédemment, nous optimisons par la suite les valeurs du paramètre k_3 dans l'objectif d'étudier la saturation des termes au niveau des requêtes : $k_1 \in [0, 100]$ avec un pas de 0,1.

Le tableau 6.4 montre les valeurs obtenues des paramètres optimisés.
Nous constatons que les valeurs présentées dans le tableau 6.4 sont toutes plus grandes que les valeurs par défaut du modèle BM25 (b et k_1).

	Avec MAP		Avec $P[0.1]$	
	b	k_1	b	k_1
$BM25S_{bin}(d,u)$, $k_3 = 0$	0,95	22	0,9	21
$BM25S_{tf}(d,u)$, $k_3 = 1000$	0,93	23	0,91	21
$BM25S_{w}(d,u)$ k_3 optimisé	0,95 $k_3 = 0,08$	22	0,9 $k_3 = 0,07$	21

Table 6.4 – Paramètres optimisés du modèle BM25S.

4.2.2 Résultats d'évaluation

En utilisant les valeurs des paramètres optimisées précédemment, nous présentons dans le tableau 6.5 les résultats d'évaluation du modèle $BM25S$ optimisé (avec des paramètres b et k_1 optimisés) par rapport au modèle de référence $BM25$ optimisé.

	MAP	$P[0.1]$
Modèle de référence $BM25(d,q)$	0,0108	0,0253
$BM25S_{bin}(d,u)$, $k_3 = 0$	0,0102	0,0213
$BM25S_{tf}(d,u)$, $k_3 = 1000$	0,0093	0,0168
$BM25S_{w}(d,u)$, k_3 optimisé	0,0116	0,0225

Table 6.5 – Résultats d'évaluation du modèle $BM25S$.

les résultats d'évaluation du modèle de RSPI $BM25S$ montrent une amélioration par rapport à un modèle classique de RI en MAP, où la meilleure valeur en MAP du modèle $BM25S$ est de 0,0116 par rapport à 0,0108 en modèle de référence BM25 ce qui nous donne une amélioration relative de 7%.
Cependant, le modèle de référence de RI classique obtient de meilleures valeurs de précision comparé au modèle de RSPI $BM25S$. Cela peut être dû au fait que le contexte de l'utilisateur contient un nombre important de termes à propos de plusieurs centres d'intérêt de cet utilisateur ce qui peut rendre la tâche plus difficile.
En ayant un contexte de l'utilisateur composé de plusieurs termes à la place de la requête de l'utilisateur contenant généralement très peu de termes (en moyenne 3 termes), nous avons voulu étudier l'impact de la saturation des termes au niveau du profil de l'utilisateur. Nous avons comparé ainsi les résultats d'évaluation des trois variantes du modèle $BM25S$ ($BM25S_{bin}$ avec une saturation maximale, $BM25S_{tf}$ sans aucune saturation et $BM25S_{w}$ avec une saturation équilibrée ou

pondérée) et nous avons constaté que parmi les trois variantes du modèle de RSPI, la variante $BM25S_w$ est celle qui donne de meilleurs résultats en MAP (0,0116) et en précision (0,0225) par rapport aux deux autres variantes $BM25S_{bin}$ et $BM25S_{tf}$.
En d'autres termes, en réduisant la saturation ($k_3 = 1000$) ou en la mettant au maximum ($k_3 = 0$), les résultats d'évaluation des variantes du modèle $BM25S$ en MAP comme en précision sont plus mauvais que le modèle de référence. Il est donc important d'équilibrer le niveau de la saturation au niveau des termes du profil social dans le cas où le profil remplace la requête de l'utilisateur.

4.3 Résultats d'évaluation du modèle $BM25S_{FreqComb}$

Nous évaluons les trois variantes du deuxième modèle de RSPI "$BM25S_{FreqComb}$" ($BM25S_{FreqComb-bin}(d,u)$, $BM25S_{FreqComb-tf}(d,u)$ et $BM25S_{FreqComb-w}(d,u)$) par rapport au modèle de référence BM25.

4.3.1 Optimisation des paramètres

De même que pour le modèle $BM25S$ (cf. section 4.2), nous optimisons de la même manière que dans la section 4.2.1, les valeurs des paramètres b, k_1 et k_3 pour le modèle $BM25S_{FreqComb}$.
De plus, comme le modèle $BM25S_{FreqComb}$ combine deux sources d'information (la requête initiale de l'utilisateur et le profil social de celui-ci), nous optimisons dans ce cas le poids w_u du profil social de l'utilisateur, intégré au sein du modèle de RSPI tel que : $w_u \in]0,1]$ avec un pas de 0,1. Le paramètre w_u nous permet d'équilibrer l'importance des termes du profil de l'utilisateur par rapport aux termes de la requête initiale.
Le tableau 6.6 montre les valeurs obtenues des paramètres optimisés :

4.3.2 Résultats d'évaluation

Les résultats d'évaluation du modèle $BM25S_{FreqComb}$ sont présentés dans le tableau 6.7.
Ces résultats renforcent notre hypothèse de RSPI qui fait que la prise en compte du profil social au niveau de la requête de l'utilisateur permet d'améliorer les résultats de la RI en MAP et en précision $P[0.1]$, où la meilleure valeur en MAP obtenue

	Avec MAP	Avec MAP		Avec $P[0.1]$	
	w_u	b	k_1	b	k_1
$BM25S_{FreqComb-bin}(d,u)$ $k_3 = 0$	0,07	0,7	18	0,7	17
$BM25S_{FreqComb-tf}(d,u)$ $k_3 = 1000$	0,04	0,8	25	0,85	21
$BM25S_{FreqComb-w}(d,u)$ $k_3 = 0,07$	0,05	0,8	25	0,85	21

Table 6.6 – Paramètres optimisés du modèle $BM25S_{FreqComb}$.

Modèles avec b et k_1 optimisés	MAP	$P[0.1]$
Modèle de référence $BM25(d,q)$	0,0108	0,0253
$BM25S_{FreqComb-bin}(d,q,u)$ $w_u = 0,07$	0,0132	0,0292
$BM25S_{FreqComb-tf}(d,q,u)$ $w_u = 0,04$	0,0119	0,0258
$BM25S_{FreqComb-w}(d,q,u)$ $w_u = 0,05$	0,0128	0,0287

Table 6.7 – Résultats d'évaluation du modèle $BM25S_{FreqComb}$.

avec le modèle $BM25S_{FreqComb}$ est de 0,0132 avec une amélioration relative de 22% par rapport au modèle de référence et la meilleure valeur en $P[0.1]$ est de 0,0292 avec une amélioration relative de 15%.

Cela veut dire que la combinaison du profil de l'utilisateur et de sa requête donne de meilleurs résultats avec les deux mesures MAP et $P[0.1]$ que le modèle classique de RI BM25 et que le modèle de RSPI $BM25S$ (MAP : 0,0132 vs 0,0116 et $P[0.1]$: 0,0292 vs 0,0225), comme montré dans le tableau 6.8.

Nous constatons aussi que toutes les variantes du modèle $BM25S_{FreqComb}$ quel que soit le niveau de la saturation considéré ($BM25S_{FreqComb-bin}$: saturation maximale, $BM25S_{FreqComb-tf}$: aucune saturation et $BM25S_{FreqComb-w}$: optimisée) donnent de meilleurs résultats par rapport au modèle de référence.

Contrairement au modèle de RSPI $BM25S$, au sein du modèle $BM25S_{FreqComb}$ la variante binaire $BM25S_{FreqComb-bin}$, où la saturation est maximale, est celle qui donne de meilleurs résultats (MAP : 0,0132) comparée aux deux autres variantes $BM25S_{FreqComb-tf}$ et $BM25S_{FreqComb-w}$.

Modèles avec b et k_1 optimisés	MAP	$P[0.1]$
$BM25S_{bin}(d,u)$, $k_3 = 0$	0,0102	0,0213
$BM25S_{tf}(d,u)$, $k_3 = 1000$	0,0093	0,0168
$BM25S_{w}(d,u)$, k_3 optimisé	0,0116	0,0225
$BM25S_{FreqComb-bin}(d,q,u)$ $w_u = 0,07$	0,0132	0,0292
$BM25S_{FreqComb-tf}(d,q,u)$ $w_u = 0,04$	0,0119	0,0258
$BM25S_{FreqComb-w}(d,q,u)$ $w_u = 0,05$	0,0128	0,0287

Table 6.8 – Résultats d'évaluation du modèle $BM25S_{FreqComb}$ vs $BM25S$.

4.4 Résultats d'évaluation du modèle $BM25S_{ScoreComb}$

Nous présentons comme pour les modèles de RSPI évalués précédemment ($BM25S$ et $BM25S_{FreqComb}$), les résultats d'évaluation des trois variantes du modèle de RSPI $BM25S_{ScoreComb}$ (binaire, fréquentielle et optimisée).

4.4.1 Optimisation des paramètres

Nous optimisons de la même manière que dans le modèle de RSPI $BM25S_{FreqComb}$ (cf. section 4.3.1), le paramètre w_u du modèle $BM25S_{ScoreComb}$ avec les valeurs optimales obtenues dans l'optimisation des paramètres du modèle de RSPI $BM25S$ (cf. section 4.2.1).

	Avec MAP	Avec MAP		Avec $P[0.1]$	
	w_u	b	k_1	b	k_1
$BM25(d,q)$		0,6	12	0,6	30
$BM25S_{ScoreComb-bin}(d,u)$ $k_3 = 0$	0,13	0,95	22	0,9	21
$BM25S_{ScoreComb-tf}(d,u)$ $k_3 = 1000$	0,3	0,93	23	0,91	21
$BM25S_{ScoreComb-w}(d,u)$ k_3 optimisé	0,26	0,95 $k_3 = 0,08$	22	0,9 $k_3 = 0,07$	21

Table 6.9 – Paramètres optimisés du modèle $BM25S_{ScoreComb}$.

	MAP	$P[0.1]$
Modèle de référence $BM25(d,q)$	0,0108	0,0253
$BM25S_{ScoreComb-bin}(d,u,q)$ $k_3 = 0$	0,0140 $w_u = 0,13$	0,0308 $w_u = 0,13$
$BM25S_{ScoreComb-tf}(d,u,q)$ $k_3 = 1000$	0,0123 $w_u = 0,3$	0,0291 $w_u = 0,58$
$BM25S_{ScoreComb-w}(d,u,q)$ k_3 optimisé	0,0137 $w_u = 0,26$	0,0306 $w_u = 0,46$

Table 6.10 – Résultats d'évaluation du modèle $BM25S_{ScoreComb}(d,u,q)$.

4.4.2 Résultats d'évaluation

Comme pour le modèle de RSPI $BM25S_{FreqComb}$ évalué précédemment, les résultats du tableau 6.10, montrent que le modèle $BM25S_{ScoreComb}$ intégrant le profil social de l'utilisateur par une combinaison de scores ($BM25(d,q)$ et $BM25S(d,u)$) donne aussi de meilleurs résultats que le modèle de référence $BM25$ et cela pour toutes les variantes du modèle $BM25S_{ScoreComb}$ (binaire : $BM25S_{ScoreComb-bin}$, fréquentielle : $BM25S_{ScoreComb-tf}$ et optimisée : $BM25S_{ScoreComb-w}$).
De même que dans le cas du modèle de RSPI précédent ($BM25S_{FreqComb}$), la variante binaire ($BM25S_{ScoreComb-bin}$) du modèle $BM25S_{ScoreComb}$ est celle qui donne de meilleurs résultats par rapport aux deux autres variantes (fréquentielle : $BM25S_{ScoreComb-tf}$ et optimisée : $BM25S_{ScoreComb-w}$).
Ainsi, nous obtenons avec une mesure MAP égale à 0,0140, une amélioration relative de 29% par rapport au modèle de référence. De même pour précision, avec une mesure de $P[0.1]$ égale à 0,0308, nous obtenons une amélioration de 21% avec la variante $BM25S_{ScoreComb-bin}$ par rapport au $BM25$.
De plus, le modèle $BM25S_{FScoreComb}$ donne de meilleurs résultats que le modèle $BM25S_{FreqComb}$ évalué précédemment en MAP (0,0140 vs 0,0132) et en précision (0,0292 vs 0,0308) comme montré dans le tableau 6.11.
Nous constatons que l'hypothèse vérifiée dans [Robertson et al., 2004] n'est pas validée avec notre collection de test dans le cas de la RSPI où le profil social de l'utilisateur est intégré par une combinaison de fréquences au lieu d'une combinaison de scores. Cette différence de comportement peut être dû à la nature des informations combinées qui correspondent à différents centres d'intérêt et de préférences de l'utilisateur et ne sont pas des descriptions spécifiques à chaque requête de l'utilisateur.

Modèles avec b et k_1 optimisés	MAP	$P[0.1]$
$BM25S_{FreqComb-bin}(d,q,u)$	0,0132 $w_u = 0,07$	0,0292
$BM25S_{FreqComb-tf}(d,q,u)$	0,0119 $w_u = 0,04$	0,0258
$BM25S_{FreqComb-w}(d,q,u)$	0,0128 $w_u = 0{,}05$	0,0287
$BM25S_{ScoreComb-bin}(d,u,q)$	0,0140 $w_u = 0,13$	0,0308 $w_u = 0,13$
$BM25S_{ScoreComb-tf}(d,u,q)$	0,0123 $w_u = 0,3$	0,0291 $w_u = 0,58$
$BM25S_{ScoreComb-w}(d,u,q)$	0,0137 $w_u = 0,26$	0,0306 $w_u = 0,46$

Table 6.11 – Résultats d'évaluation du modèle $BM25S_{ScoreComb}$ vs $BM25S_{FreqComb}$.

5 Conclusion

Dans nos expérimentations, nous avons évalué nos modèles de RSPI sur les trois collections de test que nous avons construites ($DelRSI1$, $FDelRSI1$ et $DelRSI2$). Dans ces collections de test, nous avons généré automatiquement des couples de requêtes et d'utilisateurs et des jugements de pertinence centrés utilisateur. Nous avons présenté les résultats d'évaluation de modèles de RSPI intégrant le profil social de l'utilisateur à deux niveaux : au niveau de l'indexation des documents et au niveau de la fonction de correspondance.
Nous avons montré que les modèles de RSPI proposés permettent d'améliorer les résultats de la recherche de l'utilisateur. L'étude de l'impact de la saturation au niveau des modèles de RSPI intégrant le contexte informationnel social de l'utilisateur au niveau de sa requête nous a permis de déduire les deux points suivants :

1. Dans le cas où le profil est utilisé pour remplacer la requête de l'utilisateur, une saturation optimisée donne de meilleurs résultats que le modèle classique.

2. Dans le cas d'une combinaison de la requête au profil social de l'utilisateur (combinaison de fréquences d'occurrences ou de scores) une saturation maximale donne de meilleurs résultats qu'un modèle classique en ayant optimisé

en plus, le poids du profil par rapport à la requête.

De ces résultats, nous pouvons conclure que la prise en compte du profil et du contexte social de l'utilisateur au niveau d'un modèle de RSPI peut améliorer les résultats de la recherche.
Nous constatons aussi qu'il est important de choisir le niveau de saturation des termes, qui convient dans le cas d'une utilisation du profil de l'utilisateur pour représenter un besoin d'information de l'utilisateur ou dans le cas de la combinaison du profil à la requête (combinaison de fréquence d'occurrences ou de scores). De plus, nous avons constaté que la combinaison du profil de l'utilisateur et de la requête au niveau des scores est plus intéressante qu'une combinaison au niveau des fréquences d'occurrences.

Chapitre 7

Conclusion et perspectives

Dans cette thèse, nous avons abordé une thématique récente de RI qui a vu le jour avec l'arrivée des réseaux sociaux, il s'agit de la recherche sociale personnalisée d'information.
Nous avons présenté les différentes problématiques en RSPI, conduisant à la proposition de modèles de RSPI et à l'évaluation de ces modèles, rencontrant nombreux défis notamment dans la construction d'une collection de test de RSPI dédiée.

1 État de l'art

L'étude de l'état de l'art nous a permis de catégoriser les différents travaux qui abordent la thématique de RSPI. Nous avons présenté les travaux étudiés dans trois principales catégories.
La première catégorie des travaux présentés est consacrée à l'identification des informations sociales, la seconde à l'intégration de ces informations au sein des modèles de RSPI, et la troisième, à la construction des collections de test de RSPI.

1.1 Identification des informations sociales

Dans l'identification des informations sociales issues des réseaux sociaux, nous avons recensé deux manières de considérer ces informations : comme indicateurs de l'importance sociale des entités (ressources, utilisateurs, annotations, etc.), ou pour la modélisation du profil et du contexte social de l'utilisateur. Dans cette dernière, qui correspond à notre approche d'exploitation des informations sociales, la plupart des travaux d'état de l'art utilisent les annotations et relations sociales pour construire le profil et/ou le contexte social de l'utilisateur. Différentes pondérations sont employés pour pondérer les termes au sein du contexte social de l'utilisateur (à base de TF simple, de type TF-IDF ou basée sur le modèle $BM25$).

D'autres travaux proposent d'utiliser les annotations sociales pour générer un profil social du document, appelé aussi description sociale du document à partir des annotations associées au document.

1.2 Intégration des informations sociales

L'intégration des informations sociales au sein d'un modèle de RSPI se fait à différents niveaux :

1. Au niveau de l'indexation des documents, où l'une des approches se base sur les annotations sociales associées à un document par l'utilisateur initiateur de la requête pour générer une description sociale personnalisée du document. Les auteurs calculent ainsi un score de la requête par rapport à la description personnalisée du document combiné au score thématique du document par rapport à la requête.

2. Au niveau de l'interrogation et de la fonction de correspondance. Différents travaux proposent une personnalisation à l'interrogation : en utilisant le profil de l'utilisateur à la place de la requête pour calculer un score de correspondance des documents par rapport au profil de l'utilisateur, ou en complétant la requête initiale par les termes du profil de l'utilisateur avec une combinaison au niveau des fréquences des termes ou par une combinaison du score classique des documents et descriptions des documents pour une requête et du score social des descriptions des documents pour le profil de l'utilisateur. Certains de ces travaux utilisent des pondérations simples à base de TF ou TF normalisé par la taille du contexte tandis que d'autres utilisent une pondération comme le modèle BM25, considérée comme plus appropriée qu'une pondération basique (ex : TF ou TF-IDF), au cas de documents et requêtes de tailles très variables.

   Nous avons également présenté l'état de l'art de l'utilisation des informations sociales dans les systèmes de recommandation et de filtrage collaboratif. Les résultats de ces travaux montrent l'utilité des informations sociales, qui constituent une source d'information supplémentaire à propos des utilisateurs et/ou des ressources en général.

1.3 Collection de test de RSPI

Dans la troisième catégorie de travaux de l'état de l'art étudiés, nous nous somme focalisés sur les éléments d'une collection de test de RSPI dédiée. Nous avons présenté différents travaux qui proposent d'employer les informations sociales en général et les annotations sociales en particulier, pour générer automatiquement des requêtes et des jugements de pertinence centrée utilisateur.

2 Contribution

Dans notre travail, nous avons apporté des contributions au niveau du modèle de RSPI et au niveau de la collection de test de RSPI dédiée.

2.1 Proposition de modèles de RSPI

Nous avons proposé de modéliser le contexte informationnel social de l'utilisateur à partir de ses annotations et de celles de ses relations sociales, ensuite proposé, en vue de leur exploitation en RI, deux interprétations de ces informations du contexte de l'utilisateur : une interprétation "à propos" et une interprétation "préférences". Dans notre approche, nous avons choisi l'interprétation "préférences" que nous trouvons plus adaptée au cas de la RSPI. De plus, nous avons proposé de tenir compte du contenu textuel complet du document, contrairement à la plupart des travaux de l'état de l'art dans lesquels le contenu du document est remplacé par un profil social simplifié du document.
nous avons ensuite proposé d'intégrer le contexte social de l'utilisateur au sein d'un modèle de RSPI à deux principales étapes du processus de RSPI :

1. à l'indexation personnalisée des documents : nous avons proposé un premier modèle de RSPI, appelé $BM25F_S$, intégrant le contexte informationnel social de l'utilisateur dans une indexation personnalisée des documents pour chaque utilisateur.
2. à l'interrogation, au niveau de la requête et de la fonction de correspondance et de calcul de score : nous avons proposé trois différents modèles de RSPI intégrant le contexte social de l'utilisateur.

   Le premier modèle de RSPI $BM25S$ utilise le profil social de l'utilisateur à la place de la requête.

Le second modèle de RSPI $BM25S_{FreqComb}$ permet de compléter la requête initiale de l'utilisateur en la combinant, au niveau des fréquences de termes, au profil de l'utilisateur.

Le troisième modèle de RSPI $BM25S_{ScoreComb}$ combine un score thématique classique des documents pour la requête de l'utilisateur avec un score social des documents pour le profil social de l'utilisateur.

En ayant remplacé ou complété les quelques termes de la requête (généralement ne dépassant pas trois termes) par les termes du profil de l'utilisateur (généralement composé de dizaines, voir centaines de termes), il nous a semblé important de réfléchir à la pondération appliquée à ces termes, et en particulier au niveau de saturation. Nous avons recensé trois principaux niveaux de saturation déterminé par le paramètre k_3 du modèle BM25 : saturation maximale avec un $k_3 = 0$, saturation nulle avec un $k_3 = 1000$ et une saturation optimisée avec un k_3 optimisé.

Ainsi, pour chacun de ces modèles de RSPI, nous avons proposé trois variantes correspondants aux trois niveaux de saturation cités précédemment.

2.2 Construction d'une collection de test de RSPI

Dans cette partie, nous avons proposé une approche de construction de collection de test de RSPI dédiée, avec des requêtes et jugements de pertinence centrés utilisateur. Cette approche se base sur l'utilisation des annotations sociales pour générer des requêtes par utilisateur et construire ainsi les jugements de pertinence par couple <requête, utilisateur>. Dans un premier temps, nous avons construit une première collection de test de RSPI que nous appelons $DelRSI1$, contenant approximativement 30.000 documents. Ensuite, nous avons construit une autre collection, appelée $DelRSI2$ qui est cinq fois plus grande en nombre de documents que $DelRSI1$.

3 Expérimentations

Nous avons conduit différentes expérimentations avec les trois collections de test de RSPI "$DelRSI1$", "$FDelRSI1$" et "$DelRSI2$".
Dans un premier temps, nous avons évalué le modèle classique BM25 sur "$DelRSI1$"

et "$DelRSI2$", avec deux types de données : le premier type est composé d'un ensemble de requêtes globales et de leurs jugements de pertinence et le deuxième type correspond à l'évaluation des requêtes centrées utilisateur et leurs jugements de pertinence centrée utilisateur. D'après les résultats obtenus, nous avons constaté qu'un modèle classique de RI (comme le BM25) ne semble pas adapté au cas d'une évaluation en RSPI avec des jugements de pertinence par couple de <requête, utilisateur>, d'où la nécessité d'un modèle de RSPI tenant compte des données centrées utilisateur.
Dans un second temps nous avons évalué nos différents modèles de RSPI sur les collections de test "$FDelRSI1$" et "$DelRSI2$". Nous avons constaté que dans la personnalisation à l'indexation tout comme dans la personnalisation à l'étape d'interrogations, les modèles de RSPI que nous avons proposés donnent de meilleurs résultats par rapport au modèle de référence classique $BM25$ et ce avec les deux mesures MAP et précision $P[0.1]$.
De plus, les résultats d'évaluation des modèles de RSPI visant une personnalisation des requêtes et des fonctions de correspondance et calcul de score, ont montré que comparé à une utilisation du profil de l'utilisateur pour remplacer la requête de celui-ci $BM25S$ et à une combinaison au niveau des fréquences de termes de la requête et ceux du profil $BM25S_{FreqComb}$, le modèle de RSPI $BM25S_{ScoreComb}$, correspondant à une combinaison des scores, est celui qui donne de meilleurs résultats. Nous avons aussi constaté que la saturation maximale (cas de $k_3 = 0$) est celle qui permet d'avoir de meilleurs résultats par rapport à une saturation nulle ou optimisée.

4 Perspectives

Dans nos modèles de RSPI, nous avons intégré le contexte social de l'utilisateur modélisé à partir d'un sous ensemble de données sociales, composé d'annotations sociales. Comme travaux futurs, nous comptons élargir le contexte social de l'utilisateur pour contenir d'autres ensembles et types de données. Une des perspectives liées à la modélisation du contexte social de l'utilisateur peut être dans l'étude et l'identification de différents autres types d'informations sociales à exploiter. Nous envisageons ainsi d'étudier et de proposer d'autres approches permettant d'exploiter des niveaux approfondis du voisinage social (les voisins des voisins par

exemple), qui peut se baser sur la pondération des liens entre les utilisateurs reliés à l'utilisateur initiateur de la requête.
Nous envisageons ensuite de tester différentes pondérations des données du contexte social et comparer l'utilisation du contexte social avec une interprétation "à propos" et une interprétation "préférences".
Pour ce qui est de perspectives liées aux modèles de RSPI, nous envisageons de proposer d'autres modèles de RSPI hybrides, permettant d'intégrer les informations sociales (profil de l'utilisateur) à la fois à l'indexation et à l'interrogation. Par la suite, nous trouvons important de tester d'autres modèles de pondération, autres que le BM25, pour personnaliser nos modèle de recherche sociale d'information. Nous avons évalué nos modèles de RSPI sur des collections de test de RSPI non standardisées que nous avons construites. Nous souhaitons évaluer nos modèles de RSPI sur d'autres collections de test standard conçues dans le cadre des principales compétitions de RSPI.
Enfin, dans la partie collections de test de RSPI dédiée, nous pensons tester le comportement des collections de test construites, sur différents modèles classiques et de RSPI de l'état de l'art. Ainsi, nous évaluerons nos collections de test sur différents systèmes de RSI / RSPI.
Nous envisageons aussi de mettre à disposition de la communauté scientifique en RI, les collections de test de RSPI que nous avons construites et améliorer leurs performances et probablement inclure d'autres tâches dans ces collections.

Liste de publications

Conférences nationales

1. Chahrazed Bouhini, Mathias Géry et Christine Largeron (2013). Modèle de recherche d'information centré-utilisateur. In *Proceedings of the International Conference Extraction et Gestion des Connaissances*, EGC '13, pages 275-286.
2. Chahrazed Bouhini, Mathias Géry et Christine Largeron (2012). Impact of the social networks on the information retrieval process. *Poster in W3C workshop*, WWW '12.
3. Chahrazed Bouhini (2011). Impact des Réseaux Sociaux sur le Processus de Recherche d'Information. In *COnférence en Recherche d'Information et Applications*, CORIA '11, pages 385-390.

Conférences internationales

1. Chahrazed Bouhini, Mathias Géry et Christine Largeron (2014). Integrating user's profile in the query model for Social Information Retrieval. In *Proceedings of the International Conference on Research Challenges in Information Science*, RCIS '14, pages 1-2.
2. Chahrazed Bouhini, Mathias Géry et Christine Largeron (2013). User-centered Social Information Retrieval Model Exploiting Annotations and Social Relationships. In *Proceedings of the Asia Information Retrieval Societies Conference, Information Retrieval Technology, volume 8281 of Lecture Notes in Computer Science*, AIRS '13, pages 356-367.

Bibliographie

Bibliographie

[Abel et al., 2011] Abel, F., Gao, Q., Houben, G.-J., and Tao, K. (2011). Semantic enrichment of twitter posts for user profile construction on the social web. In Antoniou, G., Grobelnik, M., Simperl, E., Parsia, B., Plexousakis, D., Leenheer, P., and Pan, J., editors, *The Semanic Web : Research and Applications*, volume 6644 of *Lecture Notes in Computer Science*, pages 375–389. Springer Berlin Heidelberg.

[Agichtein et al., 2006] Agichtein, E., Brill, E., and Dumais, S. (2006). Improving web search ranking by incorporating user behavior information. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '06, pages 19–26. ACM.

[Agrawal et al., 2009] Agrawal, R., Halverson, A., Kenthapadi, K., Mishra, N., and Tsaparas, P. (2009). Generating labels from clicks. In *Proceedings of the Second ACM International Conference on Web Search and Data Mining*, WSDM '09, pages 172–181. ACM.

[Al-Khalifa and Davis, 2006] Al-Khalifa, H. and Davis, H. (2006). Measuring the semantic value of folksonomies. In *Innovations in Information Technology*, pages 1–5.

[Allan et al., 2008] Allan, J., Aslam, J. A., Carterette, B., Pavlu, V., and Kanoulas, E. (2008). Million query track 2008 overview. In *In Proceedings of the Seventeenth Text REtrieval Conference (TREC 2007.*

[Allan et al., 2007] Allan, J., Carterette, B., Dachev, B., Aslam, J. A., Pavlu, V., and Kanoulas, E. (2007). Million query track 2007 overview. In *TREC.*

[Amati, 2003] Amati, G. (2003). *Probabilistic Models for Information Retrieval based on Divergence from Randomness.* PhD thesis, School of Computing Science, University of Glasgow.

[Amati et al., 2012] Amati, G., Amodeo, G., and Gaibisso, C. (2012). Survival analysis for freshness in microblogging search. In *21st Conference on Information and Knowledge Management*, CIKM'12, pages 2483–2486.

[Amati and Van Rijsbergen, 2002] Amati, G. and Van Rijsbergen, C. J. (2002). Probabilistic models of information retrieval based on measuring the divergence from randomness. *ACM Trans. Inf. Syst.*, 20(4) :357–389.

[Amitay et al., 2008] Amitay, E., Carmel, D., Ofek-koifman, S., Golbandi, N., and Yogev, S. (2008). Finding people and documents, using web 2.0 data. In *In Proceedings of the SIGIR 2008 Workshop on Future Challenges in Expertise Retrieval, SIGIR*, pages 1–6.

[Au-Yeung et al., 2008] Au-Yeung, C.-m., Gibbins, N., and Shadbolt, N. (2008). A study of user profile generation from folksonomies. In *Workshop on Social Web and Knowledge Management*, SWKM.

[Baeza-Yates and Ribeiro-Neto, 1999] Baeza-Yates, R. and Ribeiro-Neto, B. (1999). *Modern Information Retrieval*. Addison-Wesley.

[Bao et al., 2007] Bao, S., Xue, G., Wu, X., Yu, Y., Fei, B., and Su, Z. (2007). Optimizing web search using social annotations. In *World Wide Web*, WWW'07, pages 501–510.

[Belkin and Croft, 1992] Belkin, N. J. and Croft, W. B. (1992). Information filtering and information retrieval : two sides of the same coin ? *Commun. ACM*, 35(12) :29–38.

[Ben Jabeur et al., 2010] Ben Jabeur, L., Tamine, L., and Boughanem, M. (2010). A social model for literature access : towards a weighted social network of authors. In *9th conference Recherche d'Information Assistée par Ordinateur*, RIAO'10, pages 32–39.

[Bischoff et al., 2008] Bischoff, K., Firan, C. S., Nejdl, W., and Paiu, R. (2008). Can all tags be used for search ? In *Proceedings of the 17th ACM conference on Information and knowledge management*, CIKM '08, pages 193–202. ACM.

[Bobadilla et al., 2013] Bobadilla, J., Ortega, F., Hernando, A., and GutiéRrez, A. (2013). Recommender systems survey. *Know.-Based Syst.*, 46 :109–132.

[Bonhard and Sasse, 2006] Bonhard, P. and Sasse, M. A. (2006). 'knowing me, knowing you' – using profiles and social networking to improve recommender systems. *BT Technology Journal*, 24(3) :84–98.

[Bouadjenek et al., 2011] Bouadjenek, M. R., Hacid, H., Bouzeghoub, M., and Daigremont, J. (2011). Personalized social query expansion using social book-

marking systems. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '11, pages 1113–1114. ACM.

[Bouadjenek et al., 2013] Bouadjenek, M. R., Hacid, H., Bouzeghoub, M., and Vakali, A. (2013). Using social annotations to enhance document representation for personalized search. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '13, pages 1049–1052. ACM.

[Bouhini et al., 2013a] Bouhini, C., Géry, M., and Largeron, C. (2013a). Modèle de recherche d'information sociale centré utilisateur. In *EGC*, pages 275–286.

[Bouhini et al., 2013b] Bouhini, C., Géry, M., and Largeron, C. (2013b). User-centered social information retrieval model exploiting annotations and social relationships. In Banchs, R., Silvestri, F., Liu, T.-Y., Zhang, M., Gao, S., and Lang, J., editors, *Information Retrieval Technology*, volume 8281 of *Lecture Notes in Computer Science*, pages 356–367.

[Bouhini et al., 2014] Bouhini, C., Géry, M., and Largeron, C. (2014). Integrating user's profile in the query model for social information retrieval. In *RCIS*.

[Breuss and Tsagkias, 2014] Breuss, M. and Tsagkias, M. (2014). Learning from user interactions for recommending content in social media. In Rijke, M., Kenter, T., Vries, A., Zhai, C., Jong, F., Radinsky, K., and Hofmann, K., editors, *Advances in Information Retrieval*, volume 8416 of *Lecture Notes in Computer Science*, pages 598–604. Springer International Publishing.

[Brin and Page, 1998] Brin, S. and Page, L. (1998). The anatomy of a large-scale hypertextual web search engine. *Computer Networks*, 30(1-7) :107–117.

[Cai and Li, 2010] Cai, Y. and Li, Q. (2010). Personalized search by tag-based user profile and resource profile in collaborative tagging systems. In *19th Conference on Information and Knowledge Management*, CIKM'10, pages 969–978.

[Cai et al., 2010] Cai, Y., Li, Q., Xie, H., and Yu, L. (2010). Personalized resource search by tag-based user profile and resource profile. In *Web Information Systems Engineering*, WISE, pages 510–523.

[Carman et al., 2009] Carman, M. J., Baillie, M., Gwadera, R., and Crestani, F. (2009). A statistical comparison of tag and query logs. In *Proceedings of the*

32nd international ACM SIGIR conference on Research and development in information retrieval, SIGIR '09, pages 123–130. ACM.

[Carmel et al., 2010a] Carmel, D., Roitman, H., and Yom-Tov, E. (2010a). On the relationship between novelty and popularity of user-generated content. In *19th conference on Information and Knowledge Management*, CIKM'10, pages 1509–1512.

[Carmel et al., 2010b] Carmel, D., Roitman, H., and Yom-Tov, E. (2010b). Social bookmark weighting for search and recommendation. *The VLDB Journal*, 19(6) :761–775.

[Carmel and Yom-Tov, 2010] Carmel, D. and Yom-Tov, E. (2010). *Estimating the Query Difficulty for Information Retrieval*. Synthesis Lectures on Information Concepts, Retrieval, and Services. Morgan & Claypool Publishers.

[Carterette et al., 2008] Carterette, B., Bennett, P. N., Chickering, D. M., and Dumais, S. T. (2008). Here or there : Preference judgments for relevance. In *Proceedings of the IR Research, 30th European Conference on Advances in Information Retrieval*, ECIR'08, pages 16–27. Springer-Verlag.

[Chapelle and Zhang, 2009] Chapelle, O. and Zhang, Y. (2009). A dynamic bayesian network click model for web search ranking. In *Proceedings of the 18th International Conference on World Wide Web*, WWW '09, pages 1–10. ACM.

[Cheng et al., 2012] Cheng, C., Yang, H., King, I., and Lyu, M. (2012). Fused matrix factorization with geographical and social influence in location-based social networks. In *Twenty-Sixth AAAI Conference on Artificial Intelligence,*.

[Chirita et al., 2007] Chirita, P.-A., Firan, C. S., and Nejdl, W. (2007). Personalized query expansion for the web. In *30th Conference on Research and Development in Information Retrieval*, SIGIR'07, pages 7–14.

[Cleverdon, 1997] Cleverdon, C. (1997). Readings in information retrieval. chapter The Cranfield Tests on Index Language Devices, pages 47–59. Morgan Kaufmann Publishers Inc.

[Clinchant, 2012] Clinchant, S. (2012). *Probabilistic Models of Word Frequencies and Information Retrieval*. PhD thesis, Université Joseph Fourier de Grenoble.

[Clinchant and Gaussier, 2010] Clinchant, S. and Gaussier, É. (2010). Modèles de ri fondés sur l'information. In *CORIA*, pages 99–114.

[Croft and Harper, 1979] Croft, W. and Harper, D. (1979). Using probabilistic models of document retrieval without relevance information. *Journal of documentation*, 35(4) :285–295.

[Das et al., 2008] Das, G., Koudas, N., Papagelis, M., and Puttaswamy, S. (2008). Efficient sampling of information in social networks. In *Proceedings of the 2008 ACM Workshop on Search in Social Media*, SSM '08, pages 67–74. ACM.

[Davoodi et al., 2012] Davoodi, E., Afsharchi, M., and Kianmehr, K. (2012). A social network-based approach to expert recommendation system. In Corchado, E., Snášel, V., Abraham, A., Woźniak, M., Graña, M., and Cho, S.-B., editors, *Hybrid Artificial Intelligent Systems*, volume 7208 of *Lecture Notes in Computer Science*, pages 91–102.

[Debnath et al., 2008] Debnath, S., Ganguly, N., and Mitra, P. (2008). Feature weighting in content based recommendation system using social network analysis. In *Proceedings of the 17th international conference on World Wide Web*, WWW '08, pages 1041–1042. ACM.

[Dorigo and Caro, 1999] Dorigo, M. and Caro, G. D. (1999). *New ideas in optimization, chapter The Ant Colony Optimization Meta-Heuristic*. McGraw-Hill.

[Fang et al., 2004] Fang, H., Tao, T., and Zhai, C. (2004). A formal study of information retrieval heuristics. In *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '04, pages 49–56. ACM.

[Ferrara and Tasso, 2011] Ferrara, F. and Tasso, C. (2011). Improving collaborative filtering in social tagging systems. In *14th Conference on Advances in Artificial Intelligence : spanish association for artificial intelligence*, CAEPIA'11, pages 463–472.

[Fischer and Reuber, 1977] Fischer, E. and Reuber, A. R. (1977). The probability ranking principle in ir. *Journal of Documentation*, 33(4) :294–304.

[Fischer and Reuber, 2010] Fischer, E. and Reuber, A. R. (2010). Social interaction via new social media : (How) can interactions on twitter affect effectual thinking and behavior ? *Journal of Business Venturing*, 26 :1–18.

[Frakes, 1992] Frakes, W. B. (1992). Information retrieval. chapter Stemming algorithms, pages 131–160.

[Frakes William B., 1992] Frakes William B., B.-Y. R. (1992). *Information retrieval : data structures & algorithms*. Pearson Education.

[Géry, 2002] Géry, M. (2002). *Indexation et interrogationde chemins de lecture en contextepour la Recherche d'Information Structuréesur le Web*. PhD thesis, Université Joseph Fourier - Grenoble I.

[Géry and Largeron, 2012] Géry, M. and Largeron, C. (2012). BM25t : a BM25 extension for focused information retrieval. *Knowledge and Information Systems*, 32(1) :217–241.

[Géry et al., 2010] Géry, M., Largeron, C., and Thollard, F. (2010). une extension de bm25 pour la recherche d'information ciblee. *In Document numerique*, 13 :83–110.

[Goh and Foo, 2008] Goh, D. and Foo, S. (2008). *Social Information Retrieval Systems : Emerging Technologies and Applications for Searching the Web Effectively*. IGI Global Snippet, Hershey New York.

[Golbeck, 2006] Golbeck, J. (2006). Generating predictive movie recommendations from trust in social networks. In *Proceedings of the 4th International Conference on Trust Management*, iTrust'06, pages 93–104. Springer-Verlag.

[Gou et al., 2010] Gou, L., Zhang, X. L., Chen, H.-H., Kim, J.-H., and Giles, C. L. (2010). Social network document ranking. In *10th annual joint conference on Digital libraries*, JCDL'10, pages 313–322.

[Guy et al., 2009] Guy, I., Zwerdling, N., Carmel, D., Ronen, I., Uziel, E., Yogev, S., and Ofek-Koifman, S. (2009). Personalized recommendation of social software items based on social relations. In *Proceedings of the Third ACM Conference on Recommender Systems*, RecSys '09, pages 53–60. ACM.

[Guy et al., 2010] Guy, I., Zwerdling, N., Ronen, I., Carmel, D., and Uziel, E. (2010). Social media recommendation based on people and tags. In *33rd Conference on Research and Development in Information Retrieval*, SIGIR'10, pages 194–201.

[Halpin et al., 2007] Halpin, H., Robu, V., and Shepherd, H. (2007). The complex dynamics of collaborative tagging. In *Proceedings of the 16th international conference on World Wide Web*, WWW '07, pages 211–220. ACM.

[Hammond et al., 2005] Hammond, T., Hannay, T., Lund, B., and Scott, J. (2005). Social bookmarking tools (i) : A general overview. *D-Lib Magazine*, 11(4).

[Harman, 2000] Harman, D. (2000). What we have learned, and not learned, from trec. In *IN : BCS IRSG '2000 PROCEEDINGS*, pages 2–20.

[Harter, 1992] Harter, S. P. (1992). Psychological relevance and information science. *Journal of the American Society for Information Science*, 43(9) :602–615.

[Harvey et al., 2013] Harvey, M., Crestani, F., and Carman, M. J. (2013). Building user profiles from topic models for personalised search. In *Proceedings of the 22nd ACM international conference on Conference on information & knowledge management*, CIKM '13, pages 2309–2314. ACM.

[Heymann et al., 2008] Heymann, P., Koutrika, G., and Garcia-Molina, H. (2008). Can social bookmarking improve web search ? In *conference on Web search and web data mining*, WSDM'08, pages 195–206.

[Hotho et al., 2006] Hotho, A., Jaschke, R., Schmitz, C., and Stumme, G. (2006). Information retrieval in folksonomies : Search and ranking. In *The Semantic Web : Research and Applications*, volume 4011, pages 411–426.

[Jack et al., 2012] Jack, K., Hristakeva, M., and Garcia de Zuniga, R. (2012). Mendeley's open data for science and learning : A reply to the DataTEL challenge | mendeley. *Special issue of Datasets and Data Supported Learning in Journal of Technology Enhanced Learning*, 4 :31–46.

[Jeong et al., 2013] Jeong, J.-W., Hong, H.-K., and Lee, D.-H. (2013). i-tagranker : an efficient tag ranking system for image sharing and retrieval using the semantic relationships between tags. *Multimedia Tools Appl.*, 62(2) :451–478.

[Jimmy and Miles, 2013] Jimmy, L. and Miles, E. (2013). Overview of the trec-2013 microblog track. In *Proceedings of the Twenty-Second Text REtrieval Conference*, TREC'13.

[Joachims, 2002a] Joachims, T. (2002a). Optimizing search engines using clickthrough data. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '02, pages 133–142. ACM.

[Joachims, 2002b] Joachims, T. (2002b). Optimizing search engines using clickthrough data.

[Jones et al., 2000] Jones, K. S., Walker, S., and Robertson, S. (2000). A probabilistic model of information retrieval : development and comparative experiments, part 2. *Information Processing and Management*, 36 :809–840.

[Karweg et al., 2011] Karweg, B., Huetter, C., and Böhm, K. (2011). Evolving social search based on bookmarks and status messages from social networks. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*, CIKM '11, pages 1825–1834, New York, NY, USA. ACM.

[Kashyap et al., 2012] Kashyap, A., Amini, R., and Hristidis, V. (2012). Sonetrank : leveraging social networks to personalize search. In *CIKM*, pages 2045–2049.

[Kautz et al., 1997a] Kautz, H., Selman, B., and Shah, M. (1997a). The hidden web. *AI Magazine*, 18(2) :27–36.

[Kautz et al., 1997b] Kautz, H., Selman, B., and Shah, M. (1997b). Referral web : combining social networks and collaborative filtering. *Commun. ACM*, 40(3) :63–65.

[Khodaei and Shahabi, 2012] Khodaei, A. and Shahabi, C. (2012). Social-textual search and ranking. In *CrowdSearch*, CEUR Workshop Proceedings, pages 3–8. CEUR-WS.org.

[Kirsch, 2005] Kirsch, S. M. (2005). *Social Information Retrieval*. PhD thesis, Rheinische Friedrich-Wilhelms-Universitat Bonn.

[Konstas et al., 2009] Konstas, I., Stathopoulos, V., and Jose, J. M. (2009). On social networks and collaborative recommendation. In *32nd Conference on Research and Development in Information Retrieval*, SIGIR'09, pages 195–202.

[Koolen et al., 2012] Koolen, M., Kazai, G., Kamps, J., Doucet, A., and Landoni, M. (2012). Overview of the inex 2011 books and social search track. In Geva, S., Kamps, J., and Schenkel, R., editors, *Focused Retrieval of Content and Structure*, volume 7424 of *Lecture Notes in Computer Science*, pages 1–29. Springer Berlin Heidelberg.

[Koolen et al., 2013] Koolen, M., Kazai, G., Preminger, M., and Doucet, A. (2013). Overview of the inex 2013 social book search track. In *In CLEF 2013 Evaluation Labs and Workshop, Online Working Notes*.

[Lafferty and Zhai, 2001] Lafferty, J. and Zhai, C. (2001). Document language models, query models, and risk minimization for information retrieval. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '01, pages 111–119. ACM.

[Lagnier et al., 2013] Lagnier, C., Denoyer, L., Gaussier, E., and Gallinari, P. (2013). Predicting information diffusion in social networks using content and user's profiles. In *Proceedings of the 35th European Conference on Advances in Information Retrieval*, ECIR'13, pages 74–85. Springer-Verlag.

[Lewis and Croft, 1990] Lewis, D. D. and Croft, W. B. (1990). Term clustering of syntactic phrases. In *Proceedings of the 13th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '90, pages 385–404. ACM.

[Li and Gaussier, 2012a] Li, B. and Gaussier, É. (2012a). An information-based cross-language information retrieval model. In *ECIR*, pages 281–292.

[Li and Gaussier, 2012b] Li, B. and Gaussier, E. (2012b). Modèles d'information pour la recherche multilingue. In *COnference en Recherche d'Infomations et Applications*, CORIA'12, pages 9–24.

[Li et al., 2013] Li, L., Peng, W., Kataria, S., Sun, T., and Li, T. (2013). Frec : A novel framework of recommending users and communities in social media. In *Proceedings of the 22Nd ACM International Conference on Conference on Information & Knowledge Management*, CIKM '13, pages 1765–1770. ACM.

[Lin et al., 2011] Lin, Y., Lin, H., Jin, S., and Ye, Z. (2011). Social annotation in query expansion. In *34th Conference on Research and Development in Information Retrieval*, SIGIR'11, pages 405–414.

[Liu and Lee, 2010] Liu, F. and Lee, H. J. (2010). Use of social network information to enhance collaborative filtering performance. *Expert Systems with Applications*, 37(7) :4772 – 4778.

[Liu et al., 2013a] Liu, J., Liu, Y., Zhang, M., and Ma, S. (2013a). How do users grow up along with search engines ? : a study of long-term users' behavior.

In *Proceedings of the 22nd ACM international conference on Conference on information and knowledge management*, CIKM '13, pages 1795–1800. ACM.

[Liu et al., 2013b] Liu, X., Liu, Y., Aberer, K., and Miao, C. (2013b). Personalized point-of-interest recommendation by mining users' preference transition. In *Proceedings of the 22Nd ACM International Conference on Conference on Information and Knowledge Management*, CIKM '13, pages 733–738. ACM.

[Ma et al., 2011] Ma, H., Zhou, T. C., Lyu, M. R., and King, I. (2011). Improving recommender systems by incorporating social contextual information. *ACM Trans. Inf. Syst.*, 29(2) :9 :1–9 :23.

[Manning et al., 2008] Manning, C. D., Raghavan, P., and Schutze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press ; 1st edition.

[Markines et al., 2009] Markines, B., Cattuto, C., Menczer, F., Benz, D., Hotho, A., and Stumme, G. (2009). Evaluating similarity measures for emergent semantics of social tagging. In *World wide web*, WWW'09, pages 641–650.

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

[Mccreadie et al., 2013] Mccreadie, R., Macdonald, C., and Ounis, I. (2013). Identifying top news using crowdsourcing. *Inf. Retr.*, 16(2) :179–209.

[Mezghani et al., 2012] Mezghani, M., Zayani, C. A., Amous, I., and Gargouri, F. (2012). A user profile modelling using social annotations : a survey. In *WWW (Companion Volume)*, pages 969–976.

[Michlmayr and Cayzer, 2007] Michlmayr, E. and Cayzer, S. (2007). Learning user profiles from tagging data and leveraging them for personal(ized) information access. In *Tagging and Metadata for Social Information Organization Workshop, WWW07*.

[Mizzaro, 1997] Mizzaro, S. (1997). Relevance : The whole history. *JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE*, 48 :810–832.

[Nassr, 2002] Nassr, N. (2002). *Croisement de langues en recherche d'information : traduction et désambigu isation de requêtes*. PhD thesis, Université Paul Sabatier de Toulouse.

[Newman, 2003] Newman, M. E. J. (2003). The Structure and Function of Complex Networks. *SIAM Review*, 45(2) :167–256.

[Noll and Meinel, 2007] Noll, M.-G. and Meinel, C. (2007). Web search personalization via social bookmarking and tagging. In *The Semantic Web*, volume 4825 of *Lecture Notes in Computer Science*, pages 367–380.

[Ounis et al., 2011] Ounis, I., Craig, M., Jimmy, L., and Ian, S. (2011). Overview of the trec-2011 microblog track. In *Proceedings of twentieth Text REtrieval Conference, NIST Special Publication : SP 500-296*, TREC'11.

[Ponte and Croft, 1998] Ponte, J. M. and Croft, W. B. (1998). A language modeling approach to information retrieval. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '98, pages 275–281. ACM.

[Pujol et al., 2003] Pujol, J. M., Sangüesa, R., and Bermúdez, J. (2003). Porqpine : A distributed and collaborative search engine. In *Proceedings of the Twelfth International World Wide Web Conference - Posters*, WWW '03.

[Raiber and Kurland, 2013] Raiber, F. and Kurland, O. (2013). Using document-quality measures to predict web-search effectiveness. In *ECIR*, pages 134–145.

[Robertson et al., 1996] Robertson, S., Walker, S., Hancock-Beaulieu, M., Gatford, M., and Payne, A. (1996). Okapi at trec'4. In *The Fourth Text REtrieval Conference (TREC'4)*, TREC-4, pages 73–96.

[Robertson and Zaragoza, 2009] Robertson, S. and Zaragoza, H. (2009). The probabilistic relevance framework : Bm25 and beyond. *Found. Trends Inf. Retr.*, 3(4) :333–389.

[Robertson et al., 2004] Robertson, S., Zaragoza, H., and Taylor, M. (2004). Simple BM25 extension to multiple weighted fields. In *13th Conference on Information and Knowledge Management*, CIKM'04, pages 42–49.

[Robertson and Walker, 1994] Robertson, S. E. and Walker, S. (1994). Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In *17th Conference on Research and Development in Information Retrieval*, SIGIR'94, pages 232–241.

[Robertson and Walker, 1999] Robertson, S. E. and Walker, S. (1999). Okapi/keenbow at trec-8. In *TREC*.

[Ruthven, 2003] Ruthven, I. (2003). Re-examining the potential effectiveness of interactive query expansion. In *Proceedings of the 26th annual international*

ACM SIGIR conference on Research and development in informaion retrieval, SIGIR '03, pages 213–220. ACM.

[S.,] S., M. R. Computer and human understanding in intelligent retrieval assistance. *American Society for Information Science*, 28.

[Salton, 1969] Salton, G. (1969). A comparison between manual and automatic indexing methods. *American Documentation*, 20(1) :61–71.

[Salton, 1986] Salton, G. (1986). Another look at automatic text-retrieval systems. *Commun. ACM*, 29(7) :648–656.

[Salton and Buckley, 1988] Salton, G. and Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing Management*, 24(5) :513–523.

[Salton et al., 1983] Salton, G., Fox, E. A., and Wu, H. (1983). Extended boolean information retrieval. *Commun. ACM*, 26(11) :1022–1036.

[Sanderson, 2008] Sanderson, M. (2008). Ambiguous queries : test collections need more sense. In *31st Conference on Research and Development in Information Retrieval*, SIGIR'08, pages 499–506.

[Schamber, 1994] Schamber, L. (1994). Relevance and information behavior. *Annual Review of Information Science and Technology*, 29 :3–48.

[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.

[Shannon, 1948] Shannon, C. E. (1948). A mathematical theory of communication. *The Bell system technical journal*, 27 :379–423.

[Singhal et al., 1996] Singhal, A., Salton, G., Mitra, M., and Buckley, C. (1996). Document length normalization. *Inf. Process. Manage.*, 32(5) :619–633.

[Smith et al., 2008] Smith, M., Barash, V., Getoor, L., and Lauw, H. W. (2008). Leveraging social context for searching social media. In *Proceedings of the 2008 ACM Workshop on Search in Social Media*, SSM '08, pages 91–94. ACM.

[Soulier et al., 2012] Soulier, L., Ben Jabeur, L., Tamine, L., and Bahsoun, W. (2012). Bibrank : a language-based model for co-ranking entities in bibliographic networks. In *12th Joint Conference on Digital Libraries*, JCDL'12, pages 61–70.

[Sparck Jones, 1974] Sparck Jones, K. (1974). Automatic indexing. *J. Doc*, 30(4) :393–432.

[Stattner and Collard, 2012] Stattner, E. and Collard, M. (2012). Social-based conceptual links : Conceptual analysis applied to social networks. In *International Conference on Advances in Social Networks Analysis and Mining*, ASONAM '12, pages 25–29. IEEE Computer Society.

[Stoyanovich et al., 2008] Stoyanovich, J., Amer-Yahia, S., Marlow, C., and Yu, C. (2008). Leveraging tagging to model user interests in del.icio.us. In *AAAI Spring Symposium : Social Information Processing*, pages 104–109.

[Szomszor et al., 2008] Szomszor, M., Alani, H., Cantador, I., O'Hara, K., and Shadbolt, N. (2008). Semantic modelling of user interests based on cross-folksonomy analysis. In *Semantic Web Conference*, pages 632–648.

[Tan et al., 2005] Tan, P.-N., Steinbach, M., and Kumar, V. (2005). *Introduction to Data Mining*. Addison-Wesley ; 1st edition.

[Tan et al., 2011] Tan, S., Bu, J., Chen, C., Xu, B., Wang, C., and He, X. (2011). Using rich social media information for music recommendation via hypergraph model. *ACM Trans. Multimedia Comput. Commun. Appl.*, 7S(1) :22 :1–22 :22.

[Tchuente, 2013] Tchuente, D. (2013). *Modélisation et dérivation de profils utilisateurs à partir de réseaux sociaux : approche à partir de communautés de réseaux k-égocentriques*. PhD thesis, Université de Toulouse 3 Paul Sabatier.

[Teevan et al., 2007] Teevan, J., Dumais, S. T., and Horvitz, E. (2007). Characterizing the value of personalizing search. In *30th Conference on Research and Development in Information Retrieval*, SIGIR'07, pages 757–758.

[Teevan et al., 2011] Teevan, J., Ramage, D., and Morris, M. R. (2011). #TwitterSearch : a comparison of microblog search and web search. In *4th conference on Web search and data mining*, WSDM'11, pages 35–44.

[Theobald et al., 2005] Theobald, M., Schenkel, R., and Weikum, G. (2005). Efficient and self-tuning incremental query expansion for top-k query processing. In *28th Conference on Research and Development in Information Retrieval*, SIGIR'05, pages 242–249.

[Vallet et al., 2010] Vallet, D., Cantador, I., and Jose, J. M. (2010). Personalizing web search with folksonomy-based user and document profiles. In *32nd*

European Conference on Advances in Information Retrieval, ECIR'10, pages 420–431.

[Volkovich and Kaltenbrunner, 2011] Volkovich, Y. and Kaltenbrunner, A. (2011). Evaluation of valuable user generated content on social news web sites. In *WWW (Companion Volume)*, pages 139–140.

[Voorhees, 2005] Voorhees, Ellen M. Harman, D. K. (2005). *TREC : Experiment and Evaluation in Information Retrieval*. Digital Libraries and Electronic Publishing. The MIT Press.

[Voorhees, 2004] Voorhees, E. M. (2004). Overview of the trec 2004 robust retrieval track. In *In Proceedings of the Thirteenth Text REtrieval Conference (TREC2004*, page 13.

[Voorhees and Buckley, 2002] Voorhees, E. M. and Buckley, C. (2002). The effect of topic set size on retrieval experiment error. In *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '02, pages 316–323. ACM.

[Voorhees and Harman, 2005] Voorhees, E. M. and Harman, D. K. (2005). *TREC : Experiment and Evaluation in Information Retrieval (Digital Libraries and Electronic Publishing)*. The MIT Press.

[Wen et al., 2012] Wen, K., Li, R., Xia, J., and Gu, X. (2012). Optimizing ranking method using social annotations based on language model. *Artificial Intelligence Review*, 37 :1–16.

[Wenger, 1996] Wenger, E. (1996). How we learn. communities of practice. the social fabric of a learning organization. *J. The Healthcare Forum journal*, 39(4) :6–20.

[Wilson, 1981] Wilson, T. D. (1981). On User Studies and Information Needs. *Journal of Documentation*, 37(1) :3–15.

[Xie et al., 2012] Xie, H., Li, Q., and Cai, Y. (2012). Community-aware resource profiling for personalized search in folksonomy. *J. Comput. Sci. Technol.*, 27(3) :599–610.

[Xu et al., 2007] Xu, S., Bao, S., Cao, Y., and Yu, Y. (2007). Using social annotations to improve language model for information retrieval. In *16th Conference on Information and Knowledge Management*, CIKM'07, pages 1003—1006.

[Xu et al., 2008] Xu, S., Bao, S., Fei, B., Su, Z., and Yu, Y. (2008). Exploring folksonomy for personalized search. In *31st Conference on Research and Development in Information Retrieval*, SIGIR'08, pages 155–162.

[Yamaguchi et al., 2010] Yamaguchi, Y., Takahashi, T., Amagasa, T., and Kitagawa, H. (2010). Turank : Twitter user ranking based on user-tweet graph analysis. In Chen, L., Triantafillou, P., and Suel, T., editors, *Web Information Systems Engineering – WISE 2010*, volume 6488 of *Lecture Notes in Computer Science*, pages 240–253. Springer Berlin Heidelberg.

[Yanbe et al., 2007] Yanbe, Y., Jatowt, A., Nakamura, S., and Tanaka, K. (2007). Towards improving web search by utilizing social bookmarks. In Baresi, L., Fraternali, P., and Houben, G.-J., editors, *Web Engineering*, volume 4607, pages 343–357. Springer Heidelberg.

[Ye et al., 2011] Ye, M., Yin, P., Lee, W.-C., and Lee, D.-L. (2011). Exploiting geographical influence for collaborative point-of-interest recommendation. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '11, pages 325–334. ACM.

[Yuan et al., 2013] Yuan, Q., Cong, G., Ma, Z., Sun, A., and Thalmann, N. M. (2013). Time-aware point-of-interest recommendation. In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '13, pages 363–372. ACM.

[Zaragoza et al., 2004] Zaragoza, H., Craswell, N., Taylor, M., Saria, S., and Robertson, S. (2004). Microsoft cambridge at TREC 13 : Web and hard tracks. In *TExt Retrieval Conference*, TREC'04.

[Zhai and Lafferty, 2004] Zhai, C. and Lafferty, J. (2004). A study of smoothing methods for language models applied to information retrieval. *ACM Trans. Inf. Syst.*, 22(2) :179–214.

[Zhang et al., 2009] Zhang, X., Yang, L., Wu, X., Guo, H., Guo, Z., Bao, S., Yu, Y., and Su, Z. (2009). sDoc : exploring social wisdom for document enhancement in web mining. In *18th conference on Information and Knowledge Management*, CIKM'09, pages 395–404.