UNIVERSITE BLIDA 1 - BLIDA -

Faculté des sciences

Département d'informatique

MEMOIRE DE MAGISTER

en Informatique

Spécialité : Informatique Répartie et Mobile

INTEGRATION DE L'ANALYSE DES RESEAUX SOCIAUX

DANS LE PROCESSUS DE RECHERCHE D'INFORMATION

Par

Fatiha SADOUKI

devant le jury composé de :

H. ABED	Professeur, Université Blida 1	Présidente
A. ADLA	Professeur, Université ES-Sénia Oran	Examinateur
N. BOUSTIA	MCA, Université Blida 1	Examinatrice
N. BENBLIDIA	MCA, Université Blida 1	Promotrice
N.F. CHIKHI	MCB, Université Blida 1	Co-promoteur

Blida, Avril 2015

RESUME

L’arrivée du *Web 2.0* ou le *Web social* a remis en cause l’efficacité des techniques utilisées dans les systèmes de recherche d’information classiques qui ne répondent plus aux exigences des utilisateurs qui veulent prendre en compte leurs préférences sociales. En effet, l’utilisateur n’est plus un simple consommateur de l’information mais il participe également à sa production. Ces données générées par les utilisateurs peuvent être exploitées pour améliorer l'accès à l'information.

Dans ce travail, nous proposons d’intégrer la dimension sociale dans le processus de recherche d’information afin d’améliorer la pertinence des résultats retournés. Plus précisément, nous proposons une approche basée sur l’analyse des réseaux sociaux qui tient compte des relations sociales entre les acteurs sociaux dans le processus de RI. Nous proposons en outre un modèle de réseau d’information sociale qui représente les entités sociales qui interagissent au voisinage du document. Ce modèle inclut tous les acteurs (utilisateurs) et les données (documents et tags) qui permettent d’évaluer la pertinence sociale des documents. Cette dernière est calculée en utilisant une mesure de centralité ; elle est ensuite combinée avec la mesure de pertinence dite classique (basée uniquement sur le contenu textuel) pour calculer la pertinence globale des documents.

L’approche proposée a été évaluée expérimentalement en utilisant une collection d’articles *Wiki* dont les annotations sociales sont extraites depuis le réseau social public *Del.icio.us*. Les résultats montrent bien que la prise en compte du contexte social en recherche d'information augmente la pertinence des résultats retournés.

Mots-clés : Recherche d'information, Web 2.0, Réseau social, Social bookmarking, Mesures de centralité, Recherche d’information sociale.

ABSTRACT

The arrival of the *Web 2.0,* known also as the *social Web*, has questioned the effectiveness of the techniques used in classical information retrieval systems that no longer meet the requirements of users who want to take into account their social preferences. In fact, the user is no longer a simple consumer of information but is also involved in its production. The content generated by users could be exploited to enhance information access.

In this work, we propose to integrate social information in the information retrieval process in order to improve the information retrieval accuracy. More precisely, we propose a retrieval approach based on social network analysis. We also propose a social information network model that represents social entities that interact in the vicinity of documents. This model integrates social relationships extracted from collaborative links in the document tagging process. Moreover, we define a weighting model for social relationships which is in turn derived from the social importance of associated users. The relevance of documents is estimated by combining the document-query similarity and the document social importance derived from corresponding users.

The proposed approach has been evaluated using a collection of *Wiki* articles whose social annotations were extracted from the public social network *Del.icio.us*. Experimental results show clearly that taking into account the social context in the information retrieval process improves accuracy.

Keywords: Information retrieval, Web 2.0, Social network, Social bookmarking, Centrality measure, Social information retrieval.

الملخص

وصول الويب 2.0، المعروف أيضا باسم الشبكة الاجتماعية، شكك في فعالية التقنيات المستخدمة في نظم استرجاع المعلومات التقليدية التي لم تعد تلبي متطلبات المستخدمين الذين يرغبون أن تأخذ بعين الاعتبار التفضيلات الاجتماعية. في الواقع، أن المستخدم لم يعد مجرد مستهلك للمعلومات ولكن يشارك أيضا في إنتاجه. يمكن استغلال البيانات التي تم إنشاؤها من قبل المستخدمين لتحسين الوصول إلى المعلومات.

في هذا العمل، نقترح دمج البعد الاجتماعي في عملية البحث عن المعلومات من أجل تحسين دقة النتائج التي تم استرجاعها. بتعبير أدق، نقترح نهجا يقوم على تحليل الشبكات الاجتماعية التي تعتبر العلاقات الاجتماعية بين الفاعلين الاجتماعيين في عملية استرجاع المعلومات. كما نقترح نموذجا لشبكة المعلومات الاجتماعية التي تمثل الكيانات الاجتماعية التي تتفاعل في محيط الوثيقة. ويتضمن هذا النموذج جميع المستخدمين والبيانات (الوثائق والعلامات) التي تقيم الأهمية الاجتماعية للوثائق. وعلاوة على ذلك، نحدد نموذج الترجيح للعلاقات الاجتماعية التي بدورها مشتقة من الأهمية الاجتماعية للمستخدمين المرتبطين بها. وتقدر أهمية الوثائق من خلال الجمع بين الأهمية الكلاسيكية (تستند فقط على مضمون النص) و الأهمية الاجتماعية للوثيقة.

وقد تم تقييم النهج المقترح باستخدام مجموعة من مقالات ويكيبيديا التي تم استخراجها من الشبكة الاجتماعية العامة*Del.icio.us* . وقد اظهرت النتائج بوضوح أن الأخذ في الاعتبار السياق الاجتماعي في عملية استرجاع المعلومات يحسن الدقة.

كلمات البحث : استرجاع المعلومات، الويب 2.0، الشبكات الاجتماعية، bookmarking الاجتماعية، تدبير المركزية ، استرجاع المعلومات الاجتماعية.

REMERCIEMENTS

C'est avec un énorme plaisir que je remercie aujourd'hui tous ceux qui m'ont soutenue durant ces dernières années pour faire aboutir ce travail.

Je tiens à remercier ma promotrice, Madame Nadjia BENBLIDIA, Maître de conférences à l'université de Blida, pour m'avoir proposé ce sujet et pour avoir accepté de diriger mes travaux de recherche. Je la remercie pour la patience, la gentillesse et la disponibilité dont elle a fait preuve. Qu'elle trouve ici l'expression de ma très grande gratitude.

Je tiens aussi à exprimer ma plus profonde gratitude à mon Co-promoteur Monsieur Nacim Fateh CHIKHI, Maître de conférences à l'université de Blida, pour l'intérêt qu'il a manifesté à l'égard de mes travaux de recherche ainsi que pour son soutien et sa patience.

Je tiens également à remercier Madame Hafida ABED, Professeur à l'université de Blida1, Monsieur Abdelkader ADLA, Professeur à l'université ES-Sénia d'Oran et Madame Narhimene BOUSTIA, Maître de conférences à l'université de Blida1 pour l'intérêt qu'ils ont porté à mon travail en examinant ce mémoire et pour l'honneur qu'ils me font en participant à ce jury.

Je tiens à remercier vivement, Monsieur Mohand BOUGHANEM, Monsieur Lamjad BENJABEUR, Monsieur Arkaitz ZUBIAGA, Madame Chahrazed BOUHINI et Monsieur Ismail BADACHE pour leur aide, leur disponibilité et leur générosité pour faire avancer mon travail et mon expérimentation.

Je remercie du fond du cœur toute ma famille et belle famille, qui n'ont jamais cessé de croire en moi pendant mes années d'études et qui m'ont toujours encouragé. Merci à mon Père, qui a semé en moi l'amour du savoir.

J'adresse un merci tout particulier à mon directeur des études pour son aide et sa générosité.

Enfin, je remercie mon mari Kamel, mes amies et mes collègues qui m'ont encouragé à atteindre mes buts et soutenu dans les moments difficiles.

TABLE DES MATIERES

LISTE DES ILLUSTRATIONS, GRAPHIQUES ET TABLEAUX

INTRODUCTION

Contexte et problématique

La Recherche d'Information (RI) est un domaine qui s'intéresse à l'analyse, l'organisation, le stockage, la recherche et la découverte de l'information.

L'opération de RI est réalisée par des Systèmes de Recherche d'Information (SRI) ayant pour but de mettre en correspondance une représentation du besoin de l'utilisateur avec une représentation du contenu des documents au moyen d'une fonction de correspondance. Le défi est de trouver, parmi le volume important de documents disponibles, ceux qui correspondent au mieux à l'attente de l'utilisateur.

Ce principe de recherche est qualifié de classique, et se caractérise par un processus d'accès à l'information dépendant seulement de la disponibilité de l'information et de critères de sélection par le contenu. Le problème n'est pas tant la disponibilité de l'information, mais sa pertinence relativement à un contexte d'utilisation spécifique. Notre problématique majeure consiste à évaluer l'importance des documents.

L'avènement du Web social et l'explosion des réseaux sociaux a permis l'émergence d'une nouvelle branche de la recherche d'information : La Recherche d'Information Sociale (RIS). La RIS consiste à combiner des informations issues du contexte de la requête et du contexte des réseaux sociaux dans une même infrastructure afin d'améliorer les résultats de recherche. La prise en compte des interactions entre les utilisateurs du réseau social constitue une voie pour améliorer la performance des systèmes de recherche d'information.

Pour atteindre notre principal objectif qui consiste à améliorer le processus de recherche d'information, il est nécessaire dans un premier temps de définir le contexte social qui revient à identifier, extraire et quantifier, à partir du réseau social, une propriété sociale telle que le co-marquage ; il faut ensuite estimer une pertinence sociale intégrant cette caractéristique sociale, telle que l'importance

sociale d'autre part ; et enfin intégrer ce contexte dans le processus de recherche afin d'améliorer la qualité des documents retournés.

Contribution

Le travail présenté dans ce mémoire s'inscrit dans le domaine de la recherche d'information. Il a pour but la mise en œuvre d'une approche exploitant les réseaux sociaux en vue d'améliorer le processus de recherche d'information. Notre contribution dans ce domaine porte sur trois volets : 1) la modélisation du réseau d'information sociale, et plus précisément, l'identification des acteurs du réseau social et des relations entre eux ; 2) la prise en compte de la collaboration des utilisateurs dans l'opération de marquage des documents lors de la construction du réseau social ; et 3) l'intégration du facteur de pertinence social dans le processus de RI.

Plus précisément, nous proposons deux méthodes pour l'analyse du réseau social des utilisateurs. La première méthode s'intéresse aux utilisateurs et aux documents alors que la seconde exploite en plus les tags. Nous présentons également, deux fonctions de pondération des liens entre les utilisateurs voisins du réseau social.

Organisation du mémoire

Ce mémoire est organisé en deux parties : la première comprend deux chapitres d'état de l'art ; la seconde partie, composée des chapitres 3 et 4, présente notre contribution.

Le premier chapitre présente les concepts de base de la RI. Nous commençons par donner une définition de la RI et du processus de RI en présentant les étapes d'indexation, d'interrogation et de mise en correspondance ainsi que les techniques de reformulation des requêtes. Nous décrivons également les différents modèles servant de cadre théorique pour la modélisation du processus de RI. Ensuite, nous présentons la RI contextuelle, les différentes définitions du contexte et les possibilités de son utilisation en RI ; nous décrivons aussi les fondements théoriques et les principaux mécanismes d'évaluation des SRI. Nous présentons la

notion de pertinence, ainsi que les mesures classiques d'évaluation de cette notion. Enfin, nous présentons les campagnes d'évaluation TREC, CLEF et NTCIR.

Le deuxième chapitre aborde les thématiques de Web social et d'analyse de réseaux sociaux. Il décrit ensuite les différentes approches ainsi que la taxonomie de la recherche d'information sociale.

Le troisième chapitre expose notre approche relative à la prise en compte du réseau social pour l'amélioration du processus de RI. Nous décrivons en outre le modèle de notre réseau social.

Le quatrième chapitre présente les expérimentations réalisées afin d'évaluer notre approche. Il commence par présenter le corpus de test, puis décrit la démarche suivie pour l'évaluation des modèles proposés. L'objectif de ces expérimentations est de montrer l'intérêt de notre approche d'une part, et de valider quantitativement notre contribution d'autre part.

Nous terminons par une conclusion générale qui résume les points essentiels du travail réalisé dans le cadre de la recherche d'information sociale. Nous donnons enfin quelques perspectives de recherche quant à l'amélioration de l'approche proposée.

CHAPITRE 1
RECHERCHE D'INFORMATION : PRINCIPES, TECHNIQUES ET OUTILS

1.1 Introduction

La Recherche d'Information (RI) peut être définie comme une activité dont la finalité est de localiser et de délivrer un ensemble de documents à un utilisateur en fonction de son besoin en informations. Le défi est de pouvoir trouver, parmi le volume important de documents disponibles, ceux qui correspondent le mieux à l'attente de l'utilisateur.

L'opération de la RI est réalisée par des Systèmes de Recherche d'Information (SRI). Ces systèmes ont pour but de mettre en correspondance une représentation du besoin de l'utilisateur (requête) avec une représentation du contenu des documents au moyen d'une fonction de comparaison (ou de correspondance).

Ce chapitre est organisé en trois grandes sections : la première présente les concepts de base de la RI classique, le processus de RI, les techniques de reformulation des requêtes, et les différents modèles existants qui fournissent un cadre théorique pour la modélisation du processus de RI. La deuxième section est consacrée à la RI contextuelle. La dernière section décrit les mécanismes pour l'évaluation des SRI.

1.2 Fondements de la recherche d'information

La recherche d'information est la branche de l'informatique qui s'intéresse à l'acquisition, l'organisation, le stockage, la recherche et la distribution de l'information [1]. En bref, un système de recherche d'information permet de sélectionner à partir d'une collection de documents, des informations pertinentes répondant à des besoins utilisateurs exprimés sous forme de requêtes. Ce domaine manipule différents concepts : le besoin en information, la requête, les documents, les modèles de recherche, la pertinence, …La recherche d'information n'est pas un

domaine récent. Au début, la RI se concentrait sur les applications dans des bibliothèques. A cette époque déjà, le problème du stockage et de la recherche d'information se posait et le constat était que les volumes d'informations augmentaient, et par conséquent, les accès rapides étaient de plus en plus difficiles.

1.2.1 Définitions

Plusieurs définitions de la recherche d'information ont vu le jour dans ces dernières années ; nous citons dans ce contexte les deux définitions suivantes :

- ***Définition 1*** *:* La recherche d'information est une activité dont la finalité est de localiser et de délivrer des granules documentaires à un utilisateur en fonction de son besoin en informations [2].
- ***Définition 2*** *:* La recherche d'information est une discipline de recherche qui intègre des modèles et des techniques dont le but est de faciliter l'accès à l'information pertinente pour un utilisateur ayant un besoin en information [3].

Toutes ces définitions partagent l'idée que la RI a pour objet l'extraction à partir d'un document ou d'un ensemble de documents les informations pertinentes qui reflètent un besoin d'information.

1.2.2 Concepts de base de la RI

Plusieurs concepts clés ont été présentés dans les travaux de [4] qui sont:

a) Collection de documents *:* la collection de documents (ou fond documentaire, ou encore corpus) constitue l'ensemble des informations (documents) exploitables et accessibles.

b) Document *:* le document constitue l'information élémentaire d'une collection de documents. L'information élémentaire, appelée aussi granule de document, peut représenter tout ou une partie d'un document qui peut être retourné en réponse à une requête/ besoin d'un utilisateur.

Besoin d'information *:* la notion de besoin en information en recherche d'informations est souvent assimilée au besoin de l'utilisateur. INGWERSEN [5] défini trois types de besoin utilisateur :

- *Besoin vérificatif :* l'utilisateur cherche à vérifier le texte avec les données connues qu'il possède déjà. Il recherche donc une donnée particulière, et sait même souvent comment y accéder. La recherche d'un article sur Internet à partir d'une adresse connue serait un exemple d'un tel besoin. Un besoin de type vérificatif est dit stable, c'est-à-dire qu'il ne change pas au cours de la recherche.

- *Besoin thématique connu :* l'utilisateur cherche à clarifier, à revoir ou à trouver de nouvelles informations dans un sujet et un domaine connus. Un besoin de ce type peut être stable ou variable ; il est très possible en effet que le besoin de l'utilisateur s'affine au cours de la recherche.

- *Besoin thématique inconnu :* pour ce type de besoin, l'utilisateur cherche de nouveaux concepts ou de nouvelles relations hors des sujets ou des domaines qui lui sont familiers. Le besoin est intrinsèquement variable et est toujours exprimé de façon incomplète.

c) Requête *:* La requête formule le besoin de l'utilisateur sous forme d'un ensemble de mots exprimés en langage naturel, booléen ou graphique. La requête est soumise à un moteur de recherche pour une recherche documentaire donnée.

d) Pertinence *:* La notion de pertinence est un critère principal pour l'évaluation des systèmes de recherche d'information. La pertinence est subjective, c'est-à-dire elle dépend de l'utilisateur [6]**.**

Les études menées par BARRY [7] autour de la notion de pertinence montrent qu'elle est définie par un ensemble de critères et de préférences qui varient selon les utilisateurs (le contenu informationnel des documents, le niveau d'expertise et de connaissances de l'utilisateur, des informations liées à l'environnement, les sources des documents, ...). Ces critères sont des facteurs qui déterminent la pertinence accordée à l'information retrouvée par l'utilisateur dans un contexte de recherche précis [6].

1.2.3 Système de recherche d'information

1.2.3.1 Définition

Il existe plusieurs définitions d'un système de recherche d'information qui sont plus ou moins proches.

Selon [8] «Le but d'un système de recherche d'information est de retrouver des documents en réponse à une requête des usagers, de manière à ce que les contenus des documents soient pertinents au besoin initial d'information de l'usager».

Un système de recherche d'information est défini par un langage de représentation des documents (qui peut s'appliquer à différents corpus de documents) et des requêtes qui expriment un besoin de l'utilisateur (sous forme de mots-clés par exemple), et une fonction de mise en correspondance du besoin de l'utilisateur et du corpus de documents en vue de fournir comme résultats des documents pertinents pour l'utilisateur, c'est-à-dire répondant à son besoin d'information [9].

1.2.3.2 Processus de recherche d'information

Un SRI intègre trois fonctions principales, représentées schématiquement par le processus en U de recherche d'information [10]. La figure 1.1 illustre l'architecture générale d'un système de recherche d'information.

En plus des étapes de représentation et de recherche, quelques systèmes peuvent supporter une étape supplémentaire de reformulation automatique de requêtes. Cette étape a pour objectif d'améliorer les performances du SRI, et par conséquent d'améliorer la précision dans les réponses du système.

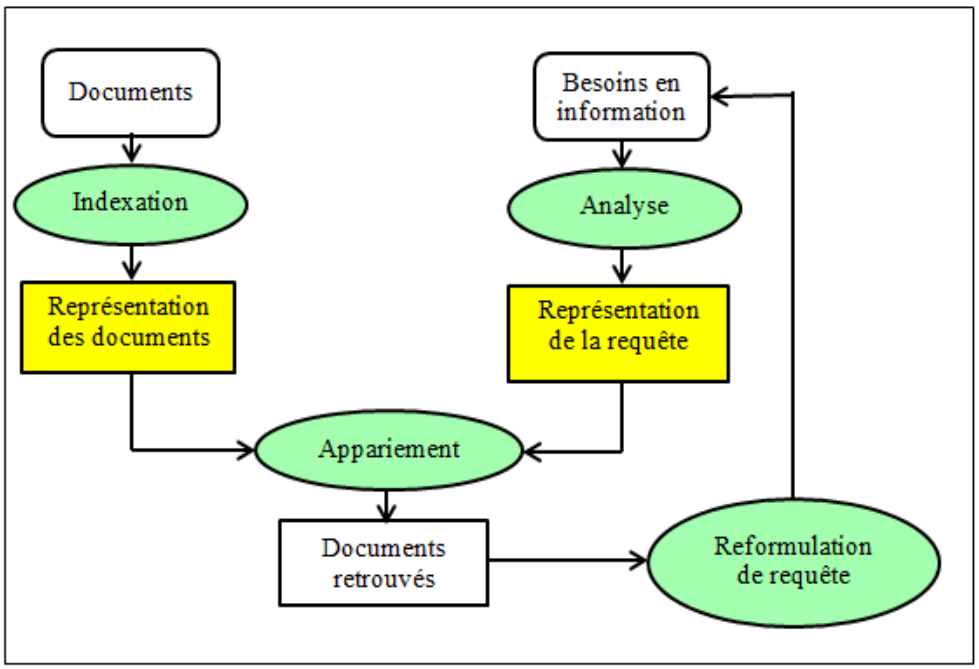

Figure 1.1 : Processus en U de recherche d'information [11]

a) Indexation / Analyse

Cette étape consiste à analyser les documents et les requêtes afin de produire un ensemble de mots clés, appelés aussi *descripteurs*, que le système pourra utiliser dans le processus de recherche ultérieur. Cette opération est appelée *indexation* [12].

Les descripteurs sont ensuite stockés dans une structure particulière appelée fichier inversé (un exemple est illustré dans la figure 1.2). Dans ce fichier, nous trouvons pour chacun de ces termes, la liste des références de chaque document le contenant.

La structure d'index est de la forme suivante (la structure de fichier inversé) : Mot → {Doc1, Doc2, …}, c'est à dire que chaque mot est associé aux documents qui le contiennent.

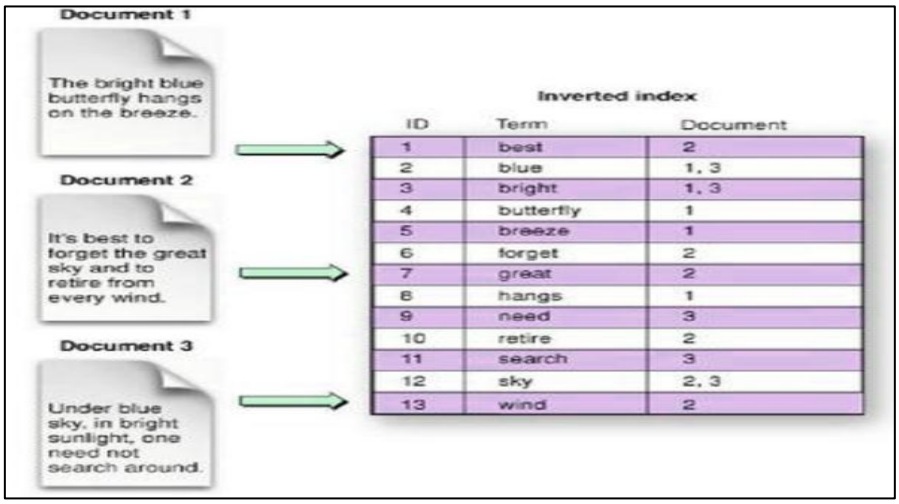

Figure 1.2 : Structure d'un index (fichier inversé)

Une requête peut être exprimée par un mot clé ou une liste de mots clés incluant des opérateurs logiques ou d'autres types d'opérateurs. Dans ce dernier cas, l'évaluation est compositionnelle, c'est à dire on commence par évaluer mot par mot ensuite on combine les listes obtenues selon l'opérateur qui relie les mots.

L'indexation peut se faire de trois manières différentes :

- **Indexation manuelle** : chaque document est analysé par un spécialiste du domaine ou un documentaliste qui va effectuer le choix des mots en utilisant un vocabulaire contrôlé (liste hiérarchique, thésaurus, lexique, . . .).
- **Indexation semi-automatique** : les termes du document sont extraits en un premier temps par un processus automatique. Puis l'indexeur, un spécialiste du domaine ou un documentaliste, intervient pour effectuer le choix final des termes significatifs et établir les relations entre les mots clés, généralement en utilisant un vocabulaire contrôlé sous forme de thésaurus ou de base terminologique.
- **Indexation automatique** : ici le processus d'indexation est complètement automatisé selon l'une des deux approches qui sont [11] : l'approche statistique qui se base sur la distribution statistique des termes dans le document, et l'approche linguistique qui se base sur les techniques de traitement du langage naturel (telles que l'analyse lexicale, syntaxique et sémantique) pour extraire les concepts les plus discriminants dans un document.

L'indexation se décompose en quatre phases :

- **L'extraction des termes (mots simples) du document** : Cette opération consiste à extraire du document un ensemble de termes ou de mots simples par une analyse lexicale permettant d'identifier les termes en reconnaissant les espaces de séparation des mots, les caractères spéciaux, les chiffres, les ponctuations, ...
- **L´élimination des mots vides** : La liste des mots simples extraite précédemment peut contenir des mots non significatifs appelés "mots vides", tels que : les pronoms personnels, les prépositions, ou encore des mots athématiques qui peuvent se retrouver dans n'importe quel document (par exemple des mots comme contenir, appartenir, …). L'élimination de ces mots peut se faire en utilisant une liste de mots vides (également appelée anti-dictionnaire), ou en écartant les mots dépassant un certain nombre d'occurrences dans la collection.
- **La normalisation** : Le processus de normalisation permet de regrouper les variantes d'un mot. En effet, on peut trouver dans un texte différentes formes d'un mot désignant le même sens. Ils seront représentés par un seul mot désignant le concept véhiculé (ex : écologie, écologiste, écologique, écologie).
- **La pondération des termes** : La pondération est une fonction fondamentale puisqu'elle traduit le degré d'importance des termes dans les documents. C'est une opération qui consiste à affecter un poids aux termes d'indexation et de recherche. Parmi les nombreuses formules de pondération définies dans le domaine, la mesure tf-idf est de loin la plus connue et la plus utilisée.

b) <u>Fonction de correspondance / Appariement</u>

Tout système de recherche d'information s'appuie sur un modèle de recherche d'information. Ce modèle se base sur une fonction de correspondance qui met en relation les termes d'un document avec ceux d'une requête en établissant une relation d'égalité entre ces termes. Cette relation d'égalité représente la base de la fonction de correspondance [9].

Le processus d'appariement est étroitement lié au processus d'indexation et de pondération des termes des requêtes et des documents du corpus.

Il existe un certain nombre de modèles théoriques dans la littérature les plus connus étant le Modèle Booléen, le Modèle Vectoriel [12] et le Modèle Probabiliste [13]. Dans le modèle booléen les requêtes sont représentées sous forme de termes reliés par des opérateurs booléens (ET, OU, NON, . . .). Le modèle vectoriel considère les documents et les requêtes comme des vecteurs pondérés, chaque élément du vecteur représentant le poids d'un terme dans la requête ou le document. Le modèle probabiliste tente d'estimer la probabilité qu'un document donné soit pertinent pour une requête donnée [14].

c) Reformulation de requêtes

L'utilisateur est souvent incapable de formuler son besoin exact en information. Par conséquent, parmi les documents qui lui sont retournés par le SRI, certains l'intéressent moins que d'autres. Compte tenu des volumes croissants des bases d'information, retrouver les informations pertinentes en utilisant seulement la requête initiale de l'utilisateur est souvent difficile. De ce fait, plusieurs techniques ont été proposées pour améliorer les performances des SRI.

La reformulation de la requête consiste à modifier la requête de l'utilisateur par ajout (ou retrait) de termes et/ou par ré-estimation de leur poids.

La reformulation peut se faire par expansion automatique de la requête ou par réinjection de pertinence. Nous présentons dans ce qui suit ces deux principales techniques :

- **Expansion automatique des requêtes (Reformulation automatique)**

L'expansion directe de la requête consiste à rajouter à la requête initiale des termes issus de ressources linguistiques existantes ou bien de ressources construites à partir des collections.

Dans ce cas, les termes sont choisis automatiquement sans l'intervention de l'utilisateur, en exploitant les premiers documents retournés par le SRI en réponse à la requête initiale [15], ou en utilisant une ressource externe qui peut être un thesaurus, une ontologie, … [16 ; 17].

- **Réinjection de pertinence (Reformulation manuelle)**

Il s'agit de la stratégie de reformulation de requêtes la plus populaire [18 ; 19]. On la nomme communément réinjection de la pertinence (ou *relevance feedback)*. Dans un cycle de réinjection de pertinence, on présente à l'utilisateur une liste de documents sélectionnés par le système comme réponse à la requête initiale. Après les avoir examinés, l'utilisateur indique ceux qu'il considère pertinents. L'idée principale de la réinjection de pertinence est de sélectionner les termes importants appartenant aux documents jugés pertinents par l'utilisateur, et de renforcer l'importance de ces termes dans la nouvelle formulation de la requête.

En fait, il s'agit de simplifier la tâche aux utilisateurs qui ne sont plus obligés de rechercher les termes importants dans les articles pertinents pour effectuer une nouvelle requête ; c'est le système qui le fait pour eux.

1.2.4 Modèles de recherche d'information

La première fonction d'un système de recherche d'information est de mesurer la pertinence d'un document vis-à-vis d'une requête. Un modèle de RI a pour rôle de fournir une formalisation du processus de recherche d'information. Il doit accomplir plusieurs rôles dont le plus important est de fournir un cadre théorique pour la modélisation de cette mesure de pertinence. Selon [20], un modèle de RI est défini par un quadruplet (D, Q, F, $R(q_i,d_j)$) : où

- D est l'ensemble des représentations des documents (vue logique) ;
- Q est l'ensemble des représentations des besoins informationnels de l'utilisateur ou ses requêtes ;
- F est le schéma du modèle théorique de représentation des documents, des requêtes et de leurs relations;
- $R(q_i,d_j)$ est la fonction de pertinence du document d_j à la requête q_i (une fonction d'ordonnancement qui associe un score de pertinence à un document).

De façon générale, les modèles de RI peuvent être classés en trois principales classes ou modèles qui sont (voir figure 1.3) :

Les modèles booléens (ensemblistes) : ces modèles trouvent leurs fondements théoriques dans la théorie des ensembles, où nous pouvons distinguer le modèle booléen pur (boolean model), le modèle booléen étendu (extended boolean model) et le modèle basé sur les ensembles flous (fuzzy set model).

Les modèles vectoriels : basés sur l'algèbre linéaire, et plus précisément le calcul vectoriel. Ils englobent le modèle vectoriel (vector model), le modèle vectoriel généralisé (generalized vector model) et la LSI (Latent Semantic Indexing).

Les modèles probabilistes : basés sur le calcul des probabilités. Ces modèles comprennent le modèle probabiliste général, le modèle de réseau de document ou d'inférence (Document Network) et le modèle de langage (ou de langue).

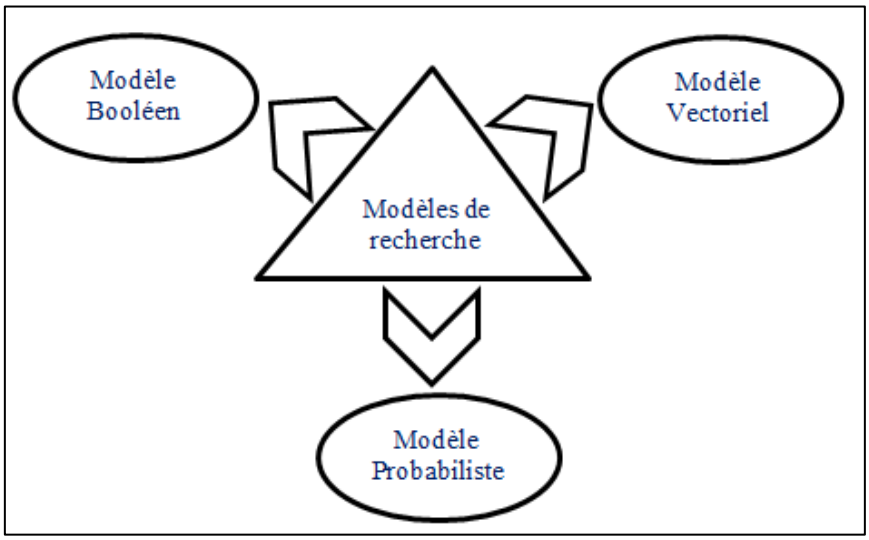

Figure 1.3 : Modèles de Recherche d'Information [21]

1.2.4.1 Modèle booléen

Le modèle booléen proposé par SALTON [12] est basé sur la théorie des ensembles. Dans ce modèle, les documents et les requêtes sont représentés par des ensembles de mots clés. Chaque document est représenté par une conjonction logique de termes non pondérés correspondant à l'index du document.

Une requête est une expression booléenne dont les termes sont reliés par des opérateurs logiques (OR, AND, NOT) permettant d'effectuer des opérations d'union, d'intersection et de différence entre les ensembles de résultats associés à chaque terme. La fonction de correspondance est basée sur l'hypothèse de

présence/absence des termes de la requête dans le document et vérifie si l'index de chaque document d_j implique l'expression logique de la requête q_i. Le résultat de cette fonction est donc binaire et est décrit comme suit : $RSV^1(q_i,d_j) = \{1,0\}$.

L'intérêt du modèle booléen est qu'il permet de faire une recherche très restrictive en fournissant à l'utilisateur une information exacte et spécifique.

Les inconvénients de ce modèle sont les suivants :

- Les réponses à une requête ne sont pas ordonnées puisque la correspondance entre un document et une requête est soit 1, soit 0. Il n'est donc pas possible de dire quel document est mieux qu'un autre. Le problème devient encore plus difficile si les documents qui répondent aux critères de la requête sont nombreux.
- L'expression d'une requête nécessite une connaissance des opérateurs booléens, ce qui n'est pas une tâche simple pour tous les usagers.

Malgré tous ces inconvénients, le modèle booléen standard reste toujours utilisé.

Différentes extensions ont été proposées pour remédier aux problèmes du modèle booléen standard. Dans ces extensions, chaque terme du document et de la requête est affecté d'un poids. Parmi ces variantes, nous citerons :

a) *Le modèle booléen étendu* [22] : dans ce modèle, la représentation de la requête reste une expression booléenne classique, tandis que les termes représentant les documents sont pondérés.

b) *Extension du modèle booléen basé sur les ensembles flous* [23]: dans ce modèle, inspiré des ensembles flous, chaque terme possède un degré d'appartenance à un document. Ce degré correspond au poids du terme dans le document.

Ces modèles ont été proposées à la fin des années 1970 et au début des années 1980. Aujourd'hui, ces extensions sont devenues standards dans le sens où la plupart des systèmes booléens utilisent un de ces modèles étendus.

---

[1] RSV : Relevance Status Value.

1.2.4.2 Modèle vectoriel

Dans les modèles vectoriels, la pertinence d'un document vis-à-vis d'une requête est définie par des mesures de distance dans un espace vectoriel [12 ; 24]. Le modèle vectoriel représente les documents et les requêtes par des vecteurs d'un espace à *M* dimensions (*M* étant le nombre de termes du vocabulaire d'indexation de la collection de documents). L'index d'un document d_j est le vecteur $W_j = (w_{1j}, w_{2j}, w_{3j}, .., w_{Mj})$, où w_{kj} dénote le poids du terme *k* dans le document d_j. Une requête est également représentée par un vecteur $W_q = (w_{1q}, w_{2q}, w_{3q}, .., w_{Mq})$, où w_{kq} est le poids du terme *k* dans la requête *q*.

Le mécanisme de recherche consiste à retrouver les vecteurs documents qui s'approchent le plus du vecteur requête. Les principales mesures de similarité utilisées sont :

Le produit scalaire :

$$RSV(q_i, d_j) = \sum_{k=1}^{M} w_{ki}.w_{kj} \tag{1.1}$$

La mesure de Jaccard :

$$RSV(q_i, d_j) = \frac{\sum_{k=1}^{M} w_{ki}.w_{kj}}{\sum_{k=1}^{M} {w_{ki}}^2 + \sum_{k=1}^{M} {w_{kj}}^2 - \sum_{k=1}^{M} w_{ki}.w_{kj}} \tag{1.2}$$

La mesure du cosinus :

$$RSV(q_i, d_j) = \frac{\sum_{k=1}^{M} w_{ki}.w_{kj}}{\left(\sum_{k=1}^{M} {w_{ki}}^2\right)^{1/2} \cdot \left(\sum_{k=1}^{M} {w_{kj}}^2\right)^{1/2}} \tag{1.3}$$

Le modèle vectoriel offre des moyens pour la prise en compte du poids des termes dans le document. Dans la littérature, plusieurs schémas de pondération ont été proposés. La majorité de ces schémas prennent en compte la pondération locale et la pondération globale.

La pondération locale permet de mesurer l'importance du terme dans le document. Elle prend en compte les informations locales du terme qui ne dépendent que du document. Elle correspond en général à une fonction de la fréquence d'occurrence du terme dans le document (noté *tf* pour *term frequency*) exprimée ainsi :

$$tf_{ij} = 1 + \log(f(t_i, d_j)) \tag{1.4}$$

où $f(t_i, d_j)$ est le nombre d'occurrences du terme t_i dans le document d_j.

Quant à la pondération globale, elle prend en compte les informations concernant le terme dans la collection. Un poids plus important doit être assigné aux termes qui apparaissent moins fréquement dans la collection car les termes qui apparaissent dans de nombreux documents de la collection ne permettent pas de distinguer les documents pertinents des documents non pertinents (on dit qu'ils sont peu utiles pour la discrimination). Un facteur de pondération globale est alors introduit. Ce facteur, appelé *idf* (*inverse document frequency*), est inversement proportionnel au nombre de documents contenant le terme. Il est calculé par :

$$idf_i = \log(\frac{N}{n_i}) \tag{1.5}$$

où n_i est le nombre de documents contenant le terme *i*, et N est le nombre total de documents de la collection.

Les fonctions de pondération combinant la pondération locale et globale sont connues sous le nom de la mesure *tf×idf.* Cette mesure donne une bonne approximation de l'importance d'un terme dans des collections de documents de tailles homogènes. Cependant, un facteur important est ignoré : la taille du document. En effet, la mesure (*tf* × *idf*) ainsi définie favorise les documents longs puisqu'ils ont tendance à répéter le même terme, ce qui accroit leur fréquence et qui augmentent par conséquence la similarité de ces documents à la requête.

Pour remédier à ce problème, des travaux ont proposé d'intégrer la taille du document dans les formules de pondération comme facteur de normalisation.

Le modèle vectoriel est l'un des modèles de RI classique les plus étudiés et les plus utilisés, à l'inverse du modèle booléen qui ne permet pas de distinguer entre deux documents qui sont indexés par les mêmes termes.

Plusieurs extensions au modèle vectoriel ont été proposées. On peut distinguer en particulier celles basées sur l'analyse sémantique latente (Latent semantic Indexing) [25] ou encore le modèle vectoriel généralisé proposé par WONG [26] qui, contrairement aux modèles classiques, ne considère pas l'hypothèse d'indépendance des termes d'indexation ce qui permet de tenir compte des dépendances qui peuvent exister entre les termes.

1.2.4.3 Modèle probabiliste

Ce modèle est fondé sur le calcul de la probabilité de pertinence d'un document pour une requête donnée [27 ; 24 ; 28]. Le principe de base consiste à retrouver des documents qui ont en même temps une forte probabilité d'être pertinents et une faible probabilité d'être non pertinents. Etant donné une requête utilisateur *Q* et un document *d,* il s'agit de calculer la probabilité de pertinence du document pour cette requête. La similarité entre un document et une requête est mesurée par le rapport entre la probabilité qu'un document *d* donné soit pertinent pour une requête *Q*, notée *p(d,Q)*, et la probabilité qu'il soit non pertinent, notée $p(\bar{d},Q)$. Cette mesure notée *RSV(Q,d)* va pouvoir classer les documents selon leurs probabilités. Plus ce score est élevé pour un document, plus ce document sera classé en haut. Il est donné par :

$$RSV(Q,d) = \frac{p(d,Q)}{p(\bar{d},Q)} \tag{1.6}$$

Une des formules les plus utilisées aujourd'hui dans le domaine de la RI est la formule *BM*25 d'OKAPI où le calcul du poids d'un terme dans un document intègre différents aspects relatifs à la fréquence locale des termes, leur rareté et la longueur des documents. Cette formule est obtenue par :

$$W(t,d) = \frac{f_{t,d} \times (k_1+1)}{k_1 \times \left((1-b) + b \times \frac{dl}{avdl}\right) + f_{t,d}} \times log\left(\frac{N - df(t,C) + 0.5}{df(t,C) + 0.5}\right) \tag{1.7}$$

avec $f_{t,d}$ est la fréquence du mot *t* dans le document *d*, *dl* est la taille du document (le nombre total d'occurrences de mots), *N* est le nombre total de documents de la collection *C*, *df(t,C)* est le nombre de documents de la collection *C* contenant *t,* k_1 et *b* sont des constantes qui dépendent des collections de test ainsi que du type des requêtes, et *avdl* représente la longueur moyenne de tous les documents dans la collection *C*.

Le modèle probabiliste est similaire au modèle vectoriel dans différents aspects, sauf qu'il trie les documents suivant leur probabilité de pertinence par rapport au besoin en information des utilisateurs au lieu d'une mesure de similarité des documents avec la requête.

1.3 Recherche d'information contextuelle

Le but fondamental de la recherche d'information contextuelle consiste à combiner des sources d'évidences issues du contexte de la requête, du contexte de l'utilisateur et de son environnement dans une même infrastructure afin de mieux caractériser les besoins en information de l'utilisateur et d'améliorer les résultats de recherche [29].

1.3.1 Définitions du contexte en RI

Les premières définitions de la notion de contexte en RI remontent aux travaux de INGERWERSEN [30] et de SARACEVIC [31] qui ont placé le contexte en amont de l'interaction utilisateur-SRI. Le contexte y est défini comme l'ensemble des facteurs cognitifs et sociaux ainsi que les buts et intentions de l'utilisateur au cours d'une session de recherche [32].

Selon [33], la RI contextuelle est définie comme suit : " *Combine search technologies and knowledge about query and user context into a single framework in order to provide the most appropriate answer for a user's information need "*.

La notion de contexte en RI peut être reliée à plusieurs entités intervenant dans le processus de recherche d'information. On distingue principalement trois entités qui interviennent dans le processus de RI [29] :

- **L'utilisateur** : il peut être caractérisé par ses connaissances, ses buts et ses intentions concernant sa recherche d'information.
- **Le système** : il s'agit des caractéristiques relatives au système-même telles que : le temps de réponse, le coût, ...
- **L'environnement de recherche** : il présente des caractéristiques liées à des critères sociaux, organisationnels et situationnels.

Parmi les éléments contextuels importants traités, nous citons le contexte social qui concerne les informations sur la communauté à laquelle appartient l'utilisateur telle que les amis, les voisins et les collègues. L'adaptation du processus de RI au contexte consiste à retourner de l'information qui répond aux préférences de la communauté des utilisateurs plutôt que de répondre aux préférences d'un seul utilisateur [34 ; 35].

1.3.2 Système de recherche d'information contextuel

1.3.2.1 Définition

Un SRI est dit contextuel ou sensible au contexte (*context-aware* en anglais) s'il exploite les données du contexte de recherche pour sélectionner l'information pertinente en réponse à une requête utilisateur [36].

1.3.2.2 Architecture d'un système de RI contextuel

Dans un système de recherche d'informations contextuel, la pertinence de l'information dépend de l'adéquation entre la requête et l'ensemble des éléments constituant le contexte qui sont perceptibles lors de la recherche [21 ; 36 ; 37]**.**

Pour pouvoir adapter les résultats de recherche au contexte de l'utilisateur, un processus de contextualisation est généralement mis en œuvre. Il consiste à construire une représentation de ces éléments contextuels. La figure 1.4 représente l'architecture d'un SRI contextuel telle qu'elle a été proposée par [36]. On peut y distinguer deux fonctionnalités fondamentales : la modélisation du contexte et l'accès contextuel à l'information.

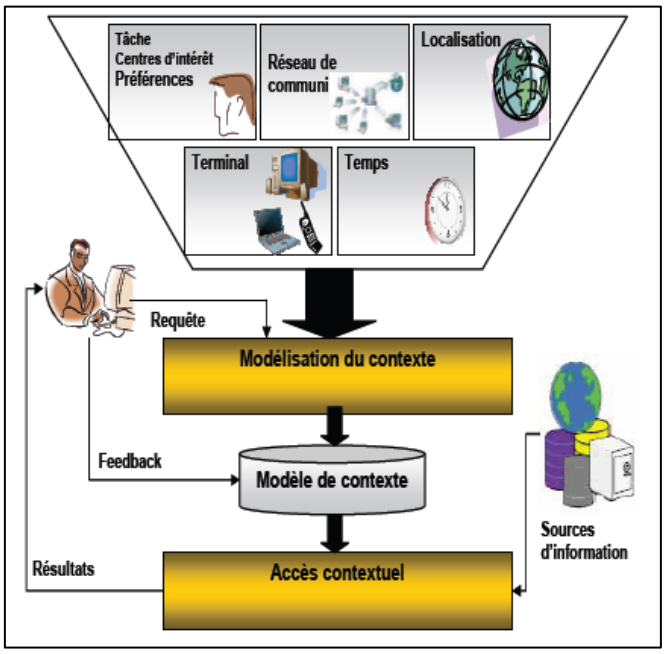

Figure 1.4 : Architecture de base d'un SRI orienté contexte [36]

a) La modélisation du contexte

Par opposition à la RI orientée système qui s'appuie sur la requête comme unique source d'évidence à modéliser, la RI contextuelle s'appuie sur une source d'évidence additionnelle exprimée à travers le contexte qu'il convient alors de modéliser. La nature et la portée du modèle dépendent des dimensions considérées du contexte. De manière générale, un modèle de contexte est défini par instanciation de chacun de ces éléments : les sources d'information exploitées, les stratégies de collecte de ces informations, les ressources de modélisation utilisées, et les modèles de représentation et d'évolution adoptés.

- **Les sources d'information :** Peuvent être de différents types : documents visités, historique des interactions, environnement (temps, température, …), ...
- **Les stratégies de collecte de ces informations :** On peut distinguer principalement deux stratégies de collecte des données du contexte : les stratégies explicites et les stratégies implicites.
  - ❖ L'acquisition explicite : repose principalement sur les techniques de feedback explicite largement utilisées dans la reformulation de requêtes par réinjection de pertinence.
  - ❖ L'acquisition implicite : consiste à collecter à l'aide d'algorithmes d'acquisition implicite les données de l'utilisateur en observant ses interactions avec le système durant les activités de recherche [38].

    L'avantage de cette approche est qu'elle ne nécessite aucune implication directe de l'utilisateur.
- **Les ressources de modélisation** : des ressources, généralement sémantiques (ontologies, dictionnaires, ...), sont parfois exploitées pour enrichir les données du modèle.
- **Les modèles de représentation** : permettent de formaliser la représentation du contexte en qualité de structure unifiée (partie d'une ontologie, classe de vecteurs de termes, ensemble de concepts, ...) ou d'un ensemble d'informations avec des structures différentes et spécifiques, puis de les faire évoluer avec le temps.

b) L'accès contextuel à l'information

C'est le processus classique de RI projeté selon une dimension additionnelle liée au contexte de recherche. Principalement, son objectif est de sélectionner l'information pertinente à la requête adressée au SRI, tenant compte de la requête d'une part et du contexte de recherche en cours d'autre part. Le contexte peut être exploité à différentes phases du processus de RI : dans la formulation de la requête, dans la fonction de pertinence, ou encore dans l'ordonnancement des résultats de recherche.

- **Reformulation de la requête**

Les éléments du contexte peuvent être utilisés pour reformuler une requête. La reformulation de requête consiste à augmenter la requête avec des informations du contexte avant de lancer le processus d'appariement.

- **Fonction d'appariement**

Le contexte peut également intervenir dans la définition de la fonction de pertinence. Le calcul du score du document est alors une fonction qui assigne au document un score de pertinence en fonction non seulement de la requête mais aussi du contexte utilisateur.

- **L'ordonnancement des résultats**

Cette phase peut également prendre en compte le contexte pour réordonner les résultats fournis par le processus de sélection. De ce fait, l'ordre final des documents à présenter à l'utilisateur est une combinaison du score/rang produit par le processus de sélection classique et celui fourni par la similitude avec le contexte de l'utilisateur.

1.4 Evaluation des systèmes de recherche d'information

L'évaluation d'un SRI consiste à mesurer ses performances vis-à-vis du besoin de l'utilisateur. A cet effet les méthodes d'évaluation largement adoptées en RI sont basées sur un modèle qui fournit une base d'évaluation comparative de l'efficacité de différents systèmes moyennant des ressources communes. Ces ressources sont essentiellement des collections de tests, des requêtes

préalablement construites, des jugements de pertinence et des métriques d'évaluation.

Le modèle d'évaluation utilisé en recherche d'information implique une collection de documents sur laquelle les recherches sont effectuées, un ensemble de requêtes de test et la liste des documents pertinents de la collection pour chacune des requêtes. Ce modèle inclut également des mesures d'évaluation permettant de contrôler l'impact sur la performance de la recherche et de la modification de certains paramètres d'un système.

1.4.1 Notion de pertinence

a) Pertinence système

Les SRI doivent s'appuyer sur un modèle de pertinence qui leur permet de calculer pour chaque document un score de pertinence. La pertinence apparaît donc ici comme une valeur numérique calculée par les SRI. Cette pertinence système a cependant des limites car elle est estimée à partir d'un score de ressemblance entre la requête et les documents, et détermine pour l'utilisateur une pertinence supposée des documents [21 ; 14].

b) Pertinence utilisateur

La pertinence utilisateur est une notion subjective permettant à ce dernier d'exprimer sa satisfaction par rapport aux documents que le système lui restitue. En effet, deux utilisateurs différents ayant soumis la même requête au SRI ne jugent pas de la même manière les réponses du système. Dans le cas où le jugement de pertinence est donné par un degré de pertinence des documents, le désaccord entre plusieurs utilisateurs est dû au fait que les besoins sont différents et que le même besoin puisse être exprimé différemment en fonction de l'utilisateur. De plus, l'interprétation que l'utilisateur fait des documents qu'il reçoit dépend en partie de ses connaissances personnelles et de son expérience, ainsi que du contexte dans lequel s'effectue sa recherche [21].

1.4.2 Evaluation des performances d'un système de recherche d'information

La performance des systèmes de recherche d'information est évaluée à partir de la pertinence des documents renvoyés. Cette notion de pertinence est ambiguë. En effet, on peut parler de pertinence objective, c'est à dire une pertinence calculée à partir des résultats du SRI (pertinence système), mais aussi de pertinence subjective : un document peut être jugé pertinent à une requête par un utilisateur (pertinence utilisateur). De même, la pertinence d'un document dépend des connaissances de l'utilisateur sur le sujet, et peut affecter la pertinence des documents examinés par la suite. C'est pour ces raisons que des mesures d'évaluation orientées utilisateurs ont été introduites.

1.4.2.1 Collection de test

Pour évaluer un SRI, on doit d'abord connaître les réponses idéales de l'utilisateur. Ainsi, l'évaluation d'un système était généralement faite avec des corpus de test. Dans un corpus de test, il y a :

- un ensemble de documents ;
- un ensemble de requêtes ;
- une liste de documents pertinents pour chaque requête.

a) *Collection de documents* : Pour qu'un corpus de test soit significatif, il faut qu'il possède un nombre de documents assez élevé. Les premiers corpus de test développés dans les années 1970 renferment quelques milliers de documents. Les corpus de test plus récents (par exemple, ceux de TREC) contiennent en général plus de 100 000 documents (considérés maintenant comme un corpus de taille moyenne), voir des millions de documents (corpus de grande taille).

b) *Requêtes* : L'évaluation d'un système ne doit pas se reposer seulement sur une requête. Pour avoir une évaluation assez objective, un ensemble de quelques dizaines de requêtes, traitant des sujets variés, est nécessaire. L'évaluation du système doit tenir compte des réponses du système pour toutes ces requêtes.

c) *Jugement de pertinence* : Finalement, il faut avoir les réponses idéales pour l'utilisateur pour chaque requête. Le dernier élément d'un corpus de test fournit cette information. Pour établir ces listes de documents pour toutes les requêtes, les

utilisateurs (ou des testeurs simulant des utilisateurs) doivent examiner chaque document de la base de test, et juger s'il est pertinent. Pour la construction d'un corpus de test, les jugements de pertinence constituent la tâche la plus difficile.

1.4.2.2 Mesures d'évaluation

L'objectif principal des systèmes de recherche d'information est d'une façon générale de retrouver tous les documents pertinents et de rejeter tous les documents non pertinents. Cet objectif est évalué par différentes mesures dont les plus courantes sont présentées ci-dessous.

a) Mesures de rappel et précision

- **Définition**

Pour mesurer la pertinence d'un SRI en terme d'efficacité, c'est-à-dire sa capacité à trouver des documents pertinents, deux (02) mesures sont largement utilisées dans la littérature : le *rappel* et la *précision*.

Précision : La précision mesure la proportion de documents pertinents retrouvés parmi tous les documents retrouvés par le système.

Rappel : Le rappel mesure la proportion de documents pertinents retrouvés parmi tous les documents pertinents dans la base.

La précision mesurée indépendamment du rappel (et inversement) est peu significative. Pour pouvoir examiner les résultats efficacement, on calcule généralement la paire de mesures (taux de rappel, taux de précision) à chaque document restitué. Les taux de *rappel* et de *précision* sont mesurés par les formules suivantes :

$$Rappel = \frac{Nombre\ de\ documents\ pertinents\ séléctionnés}{Nombre\ total\ de\ documents\ pertinents} \quad (1.8)$$

$$Précision = \frac{Nombre\ de\ documents\ pertinents\ séléctionnés}{Nombre\ total\ de\ documents\ séléctionnés} \quad (1.9)$$

Le *rappel* est défini par le nombre de documents pertinents retrouvés sur le nombre de documents pertinents de la requête. La *précision* est le nombre de documents pertinents retrouvés rapporté au nombre total de documents proposés par le moteur de recherche pour une requête donnée. Un système est dit *précis* si

peu de documents inutiles sont proposés par le système, ce qui signifie que le taux de *précision* est élevé.

Deux mesures complémentaires au rappel et à la précision, respectivement le *bruit* et le *silence* sont définies comme suit :

Bruit = 1 - Précision : donne une indication quant à la proportion de documents non pertinents renvoyés par le système.

Silence = 1 - Rappel : donne une indication quant à la proportion de documents pertinents non renvoyés par le système.

La figure 1.5 illustre la précision et le rappel d'une requête d'une façon générale. Toutefois, seule une partie des documents restituée par le système est examinée par l'utilisateur. Dans ce cas, la paire de mesures (taux de rappel, taux de précision) est calculée à chaque point de rappel (document pertinent restitué). Il s'agit de considérer la liste ordonnée des documents évalués, de calculer pour chaque document sélectionné la précision et le rappel, puis d'exprimer en fonction des valeurs trouvées la précision en fonction du rappel.

Idéalement, on voudrait qu'un système donne de bons taux de précision et de rappel en même temps. Un système qui aurait 100% pour la précision et pour le rappel signifie qu'il trouve tous les documents pertinents, et rien que les documents pertinents. Cette situation est très loin de la réalité.

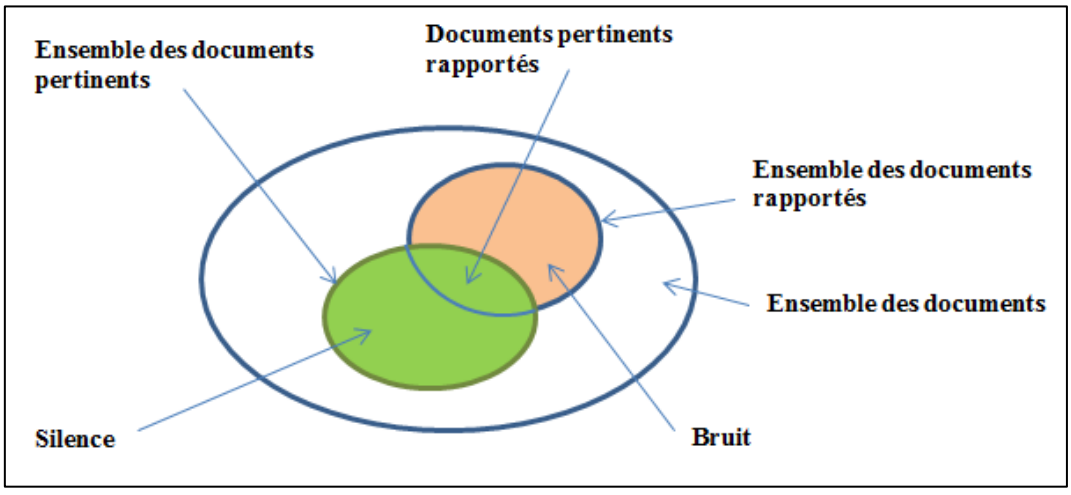

Figure 1.5 : Rappel et précision [39]

- **Courbe de rappel - précision**

Les deux métriques précédentes ne sont pas indépendantes : quand l'une augmente, l'autre diminue. On ne peut donc pas parler de la qualité d'un système en utilisant seulement une seule métrique. En effet, il est facile d'avoir 100% de rappel : il suffirait de donner toute la base comme réponse à chaque requête. Cependant, la précision dans ce cas-ci serait très basse. De même, on peut augmenter la précision en donnant très peu de documents en réponse, mais le rappel souffrira. Il faut donc utiliser les deux métriques ensemble.

Les mesures de précision-rappel ne sont pas statiques non plus (c'est-à-dire qu'un système n'a pas qu'une mesure de précision et de rappel). Le comportement d'un système peut varier en faveur de la précision ou en faveur du rappel (au détriment de l'autre métrique).

Ainsi, pour un système, on a une courbe de précision-rappel qui a en général la forme indiquée par la figure 1.6.

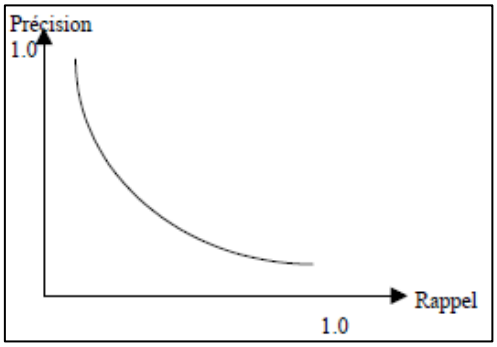

Figure 1.6 : Forme générale de la courbe Précision-Rappel d'un SRI

- **Evaluation de la Précision-Rappel**

La liste de réponses d'un système pour une requête peut varier en longueur. Une longue liste donnera ainsi un taux de rappel élevé mais un taux de précision assez bas, tandis qu'une liste courte représente le contraire. Généralement, la longueur de la liste n'est pas un paramètre inhérent d'un système ; on peut très bien le modifier selon le besoin. Mais cette modification ne doit pas altérer le comportement global du système et sa qualité. Ainsi, on peut varier cette longueur pour estimer les différents points de précision-rappel pour constituer une courbe de précision-rappel du système. Le processus d'évaluation est comme suit :

Pour i = 1, 2, … #document_dans_la_base **Faire**

Evaluer la précision et le rappel pour les i premiers documents dans la liste de réponses du système.

Fin Faire

Par exemple, soit une requête Q, et soit N = (D12, D1, D8, D3, D11, D2, D7, D210, D23, D13) l'ensemble des documents que l'on sait pertinents pour la requête Q. Soit S un SRI qui retourne les documents du tableau 1.1 en réponse à la requête Q.

[1]	D12*	[6]	D3*	[11]	D88	[16]	D31
[2]	D1*	[7]	D33	[12]	D77	[17]	D72
[3]	D17	[8]	D11*	[13]	D7*	[18]	D23*
[4]	D8*	[9]	D5	[14]	D210*	[19]	D4
[5]	D15	[10]	D2*	[15]	D18	[20]	D13*

Tableau 1.1 : Liste des documents restitués par un SRI pour la requête Q.

Dans le tableau 1.1, les documents sont ordonnés par pertinence décroissante. Les chiffres entre crochets représentent le rang du document dans la liste restituée. Les documents suivis du symbole "*" correspondent aux documents pertinents restitués par le système. Le premier document de la liste (document D12) est pertinent. On en déduit que D12 correspond à un taux de rappel de 10% (d'après la formule 1.8) puisque seulement un document pertinent sur 10 a été retrouvé à ce moment.

Lorsque le système S retrouve le document D12 au rang 1, il obtient une précision de 100% (d'après la formule 1.9) au taux de rappel de 10%.

En reprenant l'exemple présenté dans le tableau 1.1, le document pertinent suivant que S a restitué (après le document D12) est le document D1 et il se situe au rang 2. Le système obtient donc un taux de précision de 100% (2 documents pertinents sur 2 documents retournés) et un taux de rappel de 20% (2 documents pertinents retrouvés sur l'ensemble des 10 documents pertinents pour la requête).

Ce processus est continué jusqu'à l'épuisement de toute la liste de réponses du système.

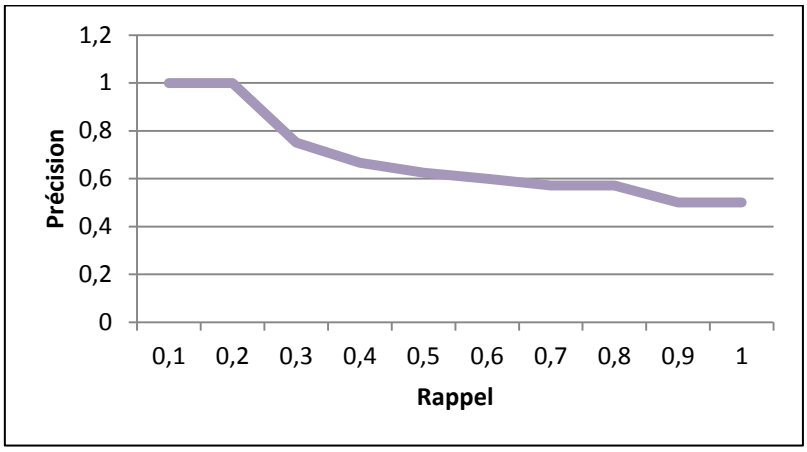

Figure 1.7 : Courbe de Rappel et Précision

La figure 1.7 illustre la courbe de rappel et précision correspondante aux résultats du tableau 1.1. Pour rendre la courbe lisible, on ne garde que la précision calculée à chaque point de rappel (c'est-à-dire à chaque document pertinent restitué).

La courbe ci-dessus permet d'évaluer les performances du système pour la requête considérée.

Afin d'évaluer le système pour un ensemble de requêtes, on calcule la moyenne des précisions à chaque niveau de rappel. Comme les niveaux de rappel ne sont pas unifiés pour l'ensemble des requêtes, on retient généralement 11 points de rappel standards de 0 à 1 avec un pas de 0,1.

Les valeurs de précision sont calculées par une interpolation linéaire. Pour deux points de rappel *i* et *j ; i* < *j,* si la précision au point *i* est inférieure à celle au point *j*, on dit que la précision interpolée à *i* égale la précision à *j.*

b) Mesures de haute précision P@X

Les mesures de haute précision permettent de ne pas évaluer l'ensemble des documents contenus dans la liste restituée par un SRI. On considère alors les X premiers documents, et la précision mesure la proportion des documents pertinents retrouvés parmi les X premiers documents retournés par le système. L'idée est qu'un système qui retourne en tête de liste un grand nombre de documents pertinents obtient une P@X supérieure à un autre système pour lequel les documents pertinents sont dispersés dans la liste restituée. Les valeurs de X peuvent être fixées à 5, 10, 15, 30, ou 100 documents par exemple. Si la valeur de X

est plus grande que le nombre total de documents retrouvés, les documents manquants sont considérés non pertinents. La valeur de haute précision correspondante est alors diminuée. Par exemple, un système qui restitue 2 documents tous pertinents aura une valeur de P@5 égale à 2/5, même si seulement 2 documents sont retrouvés.

c) R-Précision (précision exacte)

Elle correspond à la précision au point où la précision vaut le rappel. Si la requête admet R documents pertinents, la précision exacte est la précision calculée à partir des R premiers documents de la liste ordonnée des documents restitués.

Cette mesure compense les limites des mesures de haute précision quand la précision est calculée pour X documents et que le nombre total de documents |P| est inférieur à X. Si la valeur de R est plus grande que le nombre total de documents retrouvés, tous les documents non retrouvés sont alors considérés non pertinents.

Une autre mesure, dérivée de la R-précision, souvent utilisée consiste à fixer le nombre de documents retrouvés à plusieurs niveaux : top5, top10, top20, top50, ... Pour chaque niveau, on mesure la précision, et on calcule une moyenne de ces précisions sur toutes les requêtes. Cette manière de faire permet de repérer facilement les hautes précisions.

d) Précision moyenne (Mean Average Precision - MAP)

La mesure MAP est formée par une moyenne des AP (Average Precision) calculées pour chaque requête testée par le système ; chaque AP représente la moyenne des précisions calculées pour chaque document pertinent à trouver au rang de ce document. Si un document pertinent est retourné à la dixième position, la précision pour ce document est « la précision à 10 documents ». Si un document pertinent n'a pas été trouvé par le système, la précision pour ce document est nulle. Pour une requête donnée, la précision moyenne, notée AP, est calculée comme suit :

$$AP = \frac{1}{R} \sum_{i=1}^{n} p(i) \times R(i) \qquad (1.10)$$

où *R(i)* = 1 si le *i*^{ème} document restitué est pertinent, *R(i)* = 0 si le *i*^{ème} document restitué est non pertinent, *p(i)* la précision à *i* documents restitués, *R* le nombre de documents pertinents restitués et *n* le nombre total de documents retournés par le système.

Ainsi la mesure MAP est calculée pour un ensemble *C* de requêtes traité par le système comme suit :

$$MAP = \frac{1}{k} \sum_{q \in C} AP \tag{1.11}$$

avec k = |C| le nombre de requêtes dans C.

Cette mesure peut être qualifiée de globale puisqu'elle combine différents points de mesure. C'est d'ailleurs la plus souvent utilisée en RI, notamment dans le cadre des campagnes d'évaluation TREC et CLEF.

e) Mean Reciprocal Rank (MRR)

Une autre mesure basée sur le rang est la métrique Mean Reciprocal Rank. Elle permet d'évaluer le nombre de documents qu'il faut considérer avant de retrouver le premier document pertinent. Elle est égale à la moyenne, calculée sur l'ensemble des requêtes, du rang du premier document pertinent :

$$MRR = \frac{1}{|Q|} \sum_{i=1}^{|Q|} \left(\frac{1}{rank_i} \right) \tag{1.12}$$

où |Q| le nombre de requêtes de la collection.

MRR est nulle pour une requête si aucun document pertinent n'est retourné par le système. Cependant, MRR donne un score élevé pour un système qui retourne des documents pertinents en haut de la liste présentée à l'utilisateur. Cette mesure est couramment utilisée dans les systèmes Questions-Réponses où l'utilisateur s'intéresse à recevoir la bonne réponse en premier rang.

1.4.3 Campagnes d'évaluation des systèmes de recherche d'information

Pour tester l'efficacité et la performance des systèmes de recherche d'information, des campagnes d'évaluation ont été mises en place depuis les années soixante. Le projet MEDLARS (Medical Literature Analysis and Retrieval System) réalisé à la bibliothèque nationale de médecine aux Etats-Unis en est un

exemple. Les campagnes TREC (Text REtrieval Conference) créées en 1992 par le NIST (National Institut of Science and Technology) sont devenues la référence en ce qui concerne l'évaluation des systèmes. On peut aussi citer les campagnes CLEF (Cross-Language Evaluation Forum) pour l'évaluation de systèmes multilingues, les campagnes NTCIR pour les langues asiatiques, et Amaryllis pour le français.

1.4.3.1 Campagne d'évaluation TREC

La campagne d'évaluation TREC, organisée annuellement, est co-organisée par le NIST et la DARPA. Elle a pour but d'encourager la recherche documentaire basée sur de grandes collections de test, tout en fournissant l'infrastructure nécessaire pour l'évaluation des méthodologies de recherche et de filtrage d'information. Aujourd'hui, les campagnes TREC sont devenues la référence dans l'évaluation.

Les objectifs de ces campagnes étaient de favoriser l'application des techniques de RI sur de grandes collections de données, de favoriser l'échange de technologie entre les industriels et la communauté scientifique, de comparer les performances des différentes techniques, et enfin de définir un protocole d'évaluation homogène pour toute la communauté de RI.

Les participants à ces campagnes cherchent à améliorer la performance de leurs systèmes. Les conditions de participation sont les suivantes : le NIST diffuse en décembre un appel à participation qui explique les objectifs et le déroulement des tâches pour l'année à venir. Les demandes de participation doivent être déposées en Janvier. La participation à la conférence annuelle elle-même est soumise à l'envoi au NIST des résultats.

À chaque session, TREC met à la disposition des participants à la campagne un ensemble de documents et de requêtes. Pour chacune des requêtes, une liste des documents pertinents est déterminée par des juges humains. TREC met aussi à la disposition des participants un programme nommé trec-eval qui permet de calculer, pour un ensemble de requêtes, les performances des systèmes selon plusieurs critères et mesures. Plusieurs tâches sont traitées dans les campagnes TREC,

citons les principales : filtrage, recherche (ou tâche ad-hoc), interrogation inter-lingue et question-réponse.

La principale tâche est la tâche ad-hoc. Dans cette tâche, l'évaluation se fait en comparant les documents pertinents restitués par les divers systèmes participants pour une requête, et la liste des documents pertinents jugés par les juges de TREC pour la même requête testée, en utilisant le programme trec-eval. Le programme trec-eval calcule pour les 1000 premiers documents, les performances des listes restituées par les systèmes.

Dans les évaluations, une collection de 500,000 à 700,000 documents est fournie pour chaque utilisateur par le NIST qui doit l'indexer par sa propre méthode d'indexation. Avec ces documents, le NIST procure également aux participants un ensemble de 50 requêtes. Pour chaque requête, les participants classent les documents par ordre de pertinence. Les 1000 premiers documents retirés par les différents systèmes pour chaque requête sont soumis à NIST. Les examinateurs font une évaluation de pertinence des 100 ou 200 premiers documents de chaque système, et attribuent à chacun différents scores d'évaluation [40].

1.4.3.2 Autres campagnes d'évaluation

D'autres campagnes d'évaluation existent :

a) CLEF : la campagne d'évaluation CLEF a été lancée en 2000, dans le cadre d'un projet européen, pour l'évaluation des systèmes de recherches d'informations qu'ils soient monolingues ou multilingues (de langues européennes). Depuis 2002, il intègre la campagne d'évaluation des systèmes de recherches textuelles pour la langue française « Amaryllis ». Les tâches principale proposées par CLEF sont les tâches monolingue, bilingue, multilingue et les recherches dans un domaine spécifique [41].

b) NTCIR : en 1999, les campagnes NTCIR (NII-NACSIS Test Collection for IR Systems) sont apparues dans le but d'améliorer tous les domaines de l'accès à l'information y compris la recherche d'information, la production de résumés, l'extraction terminologique, ... La collection de test utilisée comprend des textes publiés entre 1998 et 1999, en chinois traditionnel, en coréen, en japonais et en anglais [6].

1.4.3.3 Limites des campagnes d'évaluation

Malgré qu'elles ont amélioré l'efficacité des systèmes, les compagnes d'évaluations ont des limites : soit au niveau du corpus de documents qui est seulement thématique et non logique et ne prend pas en compte l'opinion de l'auteur ; soit au niveau de l'évaluation faite uniquement par des professionnels ; ou encore au niveau de l'évaluation qui est faite par rapport au nombre des documents retrouvés. Or, en général un utilisateur ne cherche pas des documents mais de l'information, et la quantité d'information dépend d'un document à un autre.

1.5 Conclusion

Dans ce chapitre, nous avons présenté les principales notions et concepts de la recherche d'information classique et de la recherche d'information contextuelle. Nous avons passé en revue les méthodes et les modèles fondamentaux utilisés en RI classique, ainsi que les différentes méthodes et cadres connus d'évaluation des performances des systèmes de recherche d'information.

Dans ce chapitre, nous avons décrit le fonctionnement d'un SRI général. Dans le chapitre qui suit, nous nous intéressons à la recherche d'information sociale.

CHAPITRE 2
RECHERCHE D'INFORMATION SOCIALE

2.1 Introduction

La Recherche d'Information Sociale (RIS) est un nouveau domaine de recherche qui réunit deux domaines : la recherche d'information et l'analyse des réseaux sociaux. Elle se distingue de la RI classique par l'analyse des interactions entre les utilisateurs et de leurs profils [42]. En effet, l'analyse de ce que les gens disent, partagent et annotent permet d'identifier ce qui permettrait de mieux répondre à leurs besoins d'information [43]. Des méthodes d'analyse de réseaux sociaux sont alors appliquées pour arriver à cette fin.

2.2 Emergence du Web social

Les technologies du Web 2.0 mettent l'utilisateur au centre de la production de données et introduisent une forte composante collaborative et sociale. En conséquence, les techniques utilisées dans les systèmes de recherche d'information classiques ne répondent plus aux exigences des utilisateurs qui veulent voir leurs préférences sociales prises en compte.

Le Web 2.0 est une vision du Web mettant à disposition des utilisateurs un ensemble de services et de technologies visant à faciliter la production et le partage d'informations de manière intuitive et collaborative [44].

C'est avec l'émergence des blogs, wikis, forums et autres sites destinés à la collaboration que le Web a commencé à prendre la forme d'un outil d'intégration dans l'aspect social du Web.

De nombreux sites proposent des pages Web qui permettent à l'utilisateur de fournir ses propres textes, documents, images, vidéos, … de les incorporer au Web très facilement, sans devoir rédiger de code HTML. Ce sont des composants de ces nouveaux sites, appelés « applications Web », qui se chargent de transformer les données de l'utilisateur aux formats du Web. Parmi les nombreux types de sites sociaux nous pouvons distinguer les :

- **Blogs** : sites qui permettent à chacun d'avoir son propre journal sur le Web. Très populaire à l'époque de la naissance du Web Social, cette pratique évolue de plus en plus vers la publication de billets très courts, connue sous le nom de « microblogging ». Le blog, contrairement au wiki, met fortement l'accent sur la notion d'identité de l'auteur en tant que producteur de contenu [45].

- **flux RSS** (*Really Simple Syndication*) : Ils permettent d'envoyer à l'individu les mises à jour de divers sites qui l'intéressent (blogs, sites spécifiques, réseaux sociaux, ...) afin de l'épargner de la consultation quotidienne et la confrontation à des pages non encore mises à jour. En outre, ils sont très utiles pour faire de la veille informationnelle et recevoir tous les contenus intéressants et ce au moment de leur mise en ligne [46].

- **Wikis** : sites rendant possible la rédaction collaborative de documents. Un Wiki est un site Web dynamique et évolutif, au sens où il permet à chaque lecteur de modifier les pages consultées, d'en ajouter de nouvelles, ou encore d'en supprimer [47]. Ainsi, la dynamique d'un Wiki s'observe non seulement vis-à-vis du contenu de ses pages mais aussi via l'architecture générale de celui-ci, évoluant selon les actions des utilisateurs. Un site de type Wiki est à l'origine de Wikipedia.

- **Sites de partage de photos et vidéos** : tels que Flickr (partage de photos) et Youtube (partage de vidéos), permettent aux utilisateurs d'afficher leurs collections photographiques et leurs vidéos et de les exposer aux commentaires de la communauté en ligne [45].

- **Forums** : sites qui facilitent la discussion au sein des communautés en ligne.

- **Réseaux Sociaux** : sites destinés à la socialisation et la mise en relation des individus permettant les échanges de différents éléments (messages, photos, liens, commentaires). Ils sont de deux types : les réseaux sociaux personnels (pour

un usage personnel ou privé) et les réseaux sociaux professionnels (essentiellement pour la mise en relation avec des contacts professionnels ou des collègues). Ces deux types d'usage peuvent se trouver simultanément sur certains réseaux suite à la réduction de la frontière entre vie privée et vie professionnelle ; c'est le cas de Facebook, Myspace, LinkedIn, Viadeo, BlueKiwi, Google Plus, ... [46].

- **Outils d'indexation *(bookmarking)*** : ce sont des sites d'organisation de pages internet puisque des contenus peuvent y être marqués avec des mots clés appelés «tags» pour les classer par catégories ou par thèmes, les retrouver par la suite et les exploiter. Digg, Del.icio.us et StumbleUpon en sont des exemples [46].

Le Web social contient une grande quantité de données générées par les utilisateurs. Ces UGCs (User Generated Content) sont des statuts, des commentaires, des tags, … Ces UGC et le contenu original peuvent être exploités pour améliorer l'accès à l'information et fournir des données de qualité qui répondent aux besoins de l'utilisateur.

Comme le montre la figure 2.1, ces UGC sont utiles pour accéder à des ressources Web.

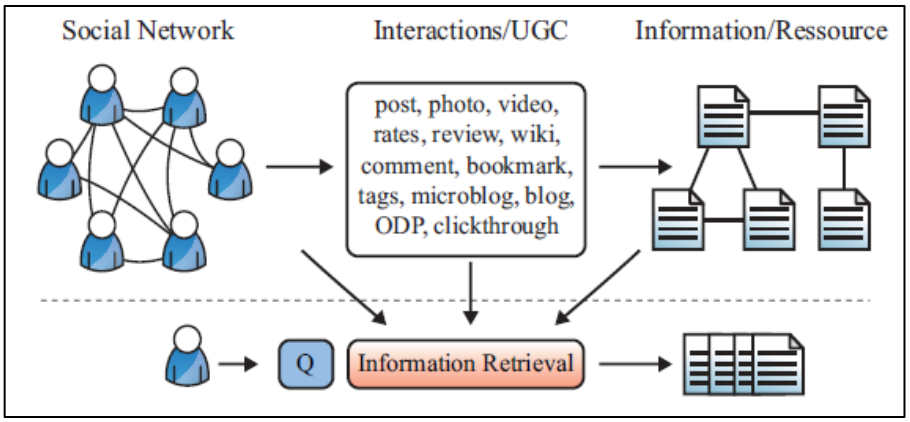

Figure 2.1 : Utilisation des UGC dans le processus de recherche d'information [43]

2.3 Analyse des réseaux sociaux

L'Analyse des Réseaux Sociaux (ARS) est définie comme étant l'étude des entités sociales (les personnes dans les organisations qu'on appelle acteurs) ainsi que leurs interactions et leurs relations. Ces interactions et relations peuvent être

représentées par un graphe ou un réseau, dans lequel chaque nœud représente un acteur et chaque lien est une relation [48].

2.3.1 Modélisation des réseaux de relations sous forme de graphes

L'application de la théorie des graphes à l'analyse des réseaux s'est incontestablement imposée. Son apport est double : d'une part, les graphes permettent une représentation graphique des réseaux de relations facilitant leur visualisation ; d'autre part, la théorie des graphes n'est pas seulement une méthode de représentation graphique mais elle présente des concepts formels permettant de qualifier, distinguer et classer les acteurs et les communautés.

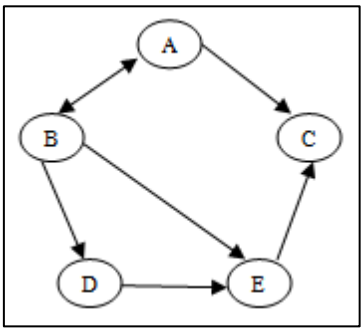

Figure 2.2 : Graphe orienté d'ordre 5

2.3.2 Représentation matricielle d'un graphe

L'idée fondamentale consiste à représenter un graphe, c'est-à-dire un ensemble de sommets et de relations (orientées ou non) entre ces sommets, par une matrice carrée, appelée « matrice d'adjacence ». Cette matrice est un tableau de chiffres qui comporte autant de colonnes que de lignes. La traduction matricielle de la figure 2.2 est la suivante :

	A	B	C	D	E
A	0	1	1	0	0
B	1	0	0	1	1
C	0	0	0	0	0
D	0	0	0	0	1
E	0	0	1	0	0

2.3.3 Réseaux sociaux

Le réseau social est un espace dans lequel les internautes interagissent (publient, partagent, annotent, commentent, ...) avec le contenu du Web. Il peut s'agir d'images (Flickr), de ressources (Twitter, Facebook, del.icio.us), ou encore d'informations professionnelles (LinkedIn). 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 et followers [49].

Un réseau social peut être représenté par un graphe G = (V,E) où l'ensemble des nœuds V représente les entités sociales (les acteurs) et l'ensemble des arcs E = (V × V) représente les relations sociales entre eux.

Les réseaux sociaux nous aident à comprendre le rôle de l'ensemble des acteurs et leurs relations[2]. L'analyse de ces réseaux se fait avec un vocabulaire plutôt formel et abstrait emprunté également à la théorie mathématique des graphes.

2.3.4 Mesures de centralité issues de l'analyse des réseaux sociaux

La centralité a été traitée dans différents domaines : i) Dans le domaine des recherches sur le web pour développer des algorithmes de classification de thématiques, où les pages et les hyperliens vers les pages Web sont représentés comme des nœuds et des arêtes dans un graphe [50] ; ii) Dans les réseaux sociaux, où la centralité est utilisée pour représenter l'acteur principal qui travaille sur un sujet central [51].

La centralité est une caractéristique de la position (la popularité ou la visibilité) d'un nœud dans un réseau. Les acteurs importants sont ceux qui sont fortement liés et impliqués avec les autres acteurs. Dans le cadre d'une organisation, une personne qui a beaucoup de contacts et qui communique souvent avec les autres personnes est considérée plus importante qu'une autre personne ayant moins de contacts. Ces contacts sont modélisés par des liens. Un acteur central est un acteur

[2] La relation ou lien qui unit deux acteurs correspond à l'ensemble des interactions existantes entre eux. Au-delà de la simple interaction, elle porte une valeur (amitié, liens de parenté, liens hiérarchiques, contacts professionnels, liens de voisinage,…) ou un ensemble de valeurs.

qui est impliqué dans plusieurs liens. La figure 2.3 montre un exemple simple qui utilise un graphe non dirigé. Chaque nœud dans le réseau social est un acteur et chaque lien indique que les deux acteurs aux extrémités communiquent ensemble. Intuitivement, nous remarquons que l'acteur 1 est l'acteur le plus central parce qu'il communique avec la majorité des autres acteurs. Il y a plusieurs types de centralités, mais les plus citées sont la centralité de degré, la centralité de proximité et la centralité d'intermédiarité.

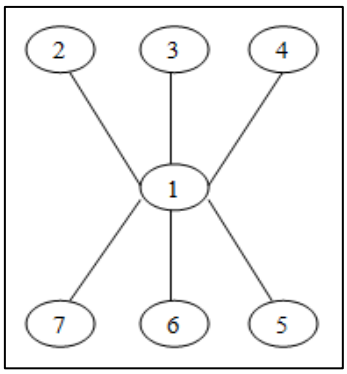

Figure 2.3 : Exemple de graphe non dirigé

2.3.4.1 Centralité de degré

Le degré est peut-être la mesure la plus simple à juger quant à la centralité d'un acteur. On l'obtient en calculant le nombre total (i.e. la somme) de connexions directes qu'a un acteur avec les autres membres du réseau [52]. Il peut être calculé en chiffres absolus (dans ce cas sa magnitude va dépendre de la taille du réseau) ou être standardisé, en divisant le nombre de liens directs d'un sommet par le nombre maximum de liens possibles.

Le degré permet de démontrer l'intégration ou l'isolement d'un acteur dans l'ensemble du réseau [53]. Sa définition dépend du type de graphe :

a) Graphe non dirigé : Dans un graphe non dirigé, le degré de centralité d'un acteur v_i est le degré du nœud acteur (le nombre d'arêtes) normalisé par le degré maximal *(N – 1)* et est défini par :

$$C^{deg}(v_i) = \frac{1}{N-1} \sum_{j=1}^{N} a_{ij} \qquad (2.1)$$

où *N* le nombre total d'acteurs dans le réseau et a_{ij} représente le lien entre l'acteur *i* avec l'acteur *j*.

b) Graphe dirigé : Dans ce cas, nous distinguons entre les liens entrants d'un acteur v_i (les liens pointant vers v_i) et les liens sortants (les liens pointant à partir de v_i). Le degré de centralité est défini par deux mesures ; une par rapport aux liens sortants et une par rapport aux liens entrants :

$$C_{out}^{deg}(v_i) = \frac{1}{N-1} \sum_{j=1}^{N} a_{ij} \qquad C_{in}^{deg}(v_i) = \frac{1}{N-1} \sum_{j=1}^{N} a_{ji}$$

La figure 2.4 indique la centralité de degré pour les nœuds du graphe dirigé G suivant [54] :

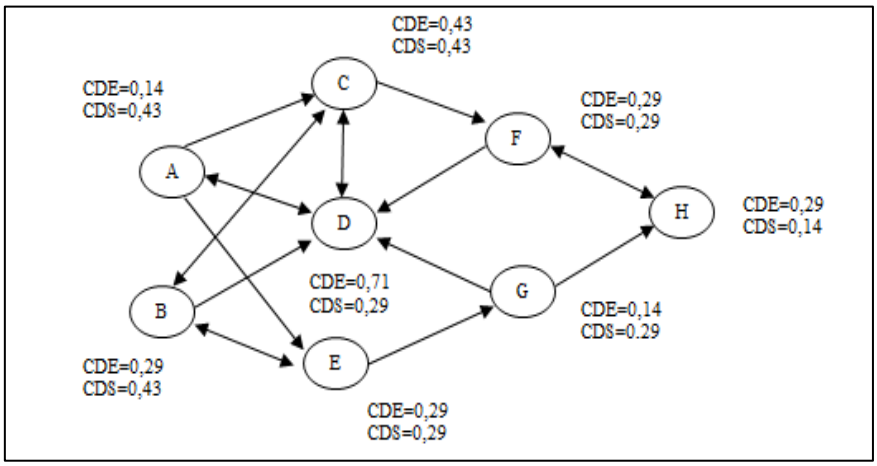

Figure 2.4 : Centralités de degré entrant (CDE) et sortant (CDS)

Les résultats montrent que les nœuds A, B et C possèdent la plus forte centralité par rapport aux liens sortants, tandis que le nœud D possède la plus forte centralité par rapport aux liens entrants.

Les acteurs centraux sont les acteurs les plus actifs et qui ont le plus de liens avec les autres acteurs

2.3.4.2 Centralité de proximité

La centralité de proximité pour un sommet dépend inversement de la somme des chemins minimaux entre ce sommet et tous les autres [55]. Cette approche définit la centralité en utilisant la notion de proximité ou de distance.

Cette mesure correspond à l'idée qu'un acteur est important (central) s'il est capable de contacter facilement un grand nombre d'acteurs avec un minimum d'effort (l'effort ici est relatif à la taille des chemins). Par conséquence, sa distance avec les autres doit être courte.

a) Graphe non dirigé **:** La centralité $C_c(i)$ de proximité d'un acteur v_i est définie par :

$$C_c(v_i) = \frac{N-1}{\sum_{j=1}^{n} dist(v_i,v_j)} \tag{2.2}$$

où N le nombre total d'acteurs dans le réseau et $dist(v_i,v_j)$ est la distance la plus courte à partir de l'acteur v_i vers l'acteur v_j (mesurée par le nombre de liens via le chemin le plus court).

La valeur de cette mesure varie entre 0 et 1 comme (N – 1) est la valeur minimale du dénominateur, qui correspond à la somme de la distance la plus courte à partir de l'acteur v_i vers les autres acteurs. Notons bien que cette équation est valable seulement dans un graphe connecté.

b) Graphe dirigé : La même équation est utilisée pour les graphes dirigés à la différence que le calcul de la distance doit intégrer la direction des liens.

La centralité de proximité est définie par deux mesures : une par rapport aux liens sortants et une par rapport aux liens entrants :

$$C_{out}^{pro}(v_i) = \frac{N-1}{\sum_{j=1}^{n} dist(v_i,v_j)} \qquad C_{in}^{pro}(v_i) = \frac{N-1}{\sum_{j=1}^{n} dist(v_j,v_i)}$$

Pour le calcul des distances entre sommets, Freeman propose d'utiliser la distance géodésique (i.e. taille du chemin le plus court) entre les nœuds. La figure 2.5 indique la centralité de proximité pour les nœuds du graphe G.

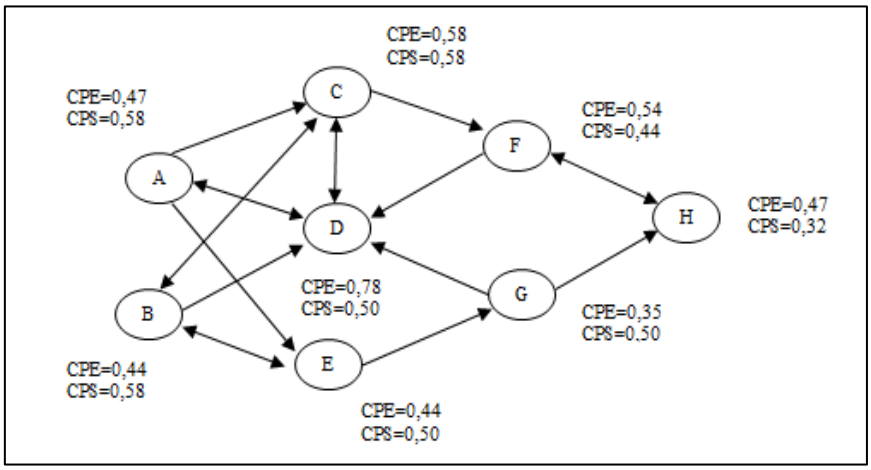

Figure 2.5 : Centralités de proximité entrante (CPE) et sortante (CPS)

Les nœuds A, B et C possèdent la plus forte centralité par rapport aux liens sortants, tandis que le nœud D possède la plus forte centralité par rapport aux liens entrants.

Le nœud le plus central est le moins loin en moyenne des autres.

2.3.4.3 Centralité d'intermédiarité

La centralité d'intermédiarité pour un sommet dépend directement du nombre des chemins minimaux qui passent par ce sommet.

Si deux acteurs non adjacents v_j et v_k veulent communiquer et si l'acteur v_i se localise sur le chemin entre v_j et v_k, alors v_i a un certain contrôle sur leur interaction. L'intermédiarité mesure ce contrôle de v_i sur les deux autres acteurs. Par conséquent, si v_i se localise sur le chemin de plusieurs interactions alors v_i est un acteur important.

a) Graphe non dirigé : Soit g_{jk} le nombre des chemins les plus courts entre les deux acteurs v_j et v_k. L'intermédiarité d'un acteur v_i est définie par le nombre des chemins les plus courts entre v_j et v_k passant par v_i notés par $g_{jk}(v_i)$ (avec $v_j \neq v_i$ et $v_k \neq v_i$), normalisé par le nombre total des chemins les plus courts entre toutes les paires d'acteurs qui n'incluent pas v_i :

$$C^{int}(v_i) = \sum_{j=1}^{N} \sum_{k=1}^{N} \left(\frac{g_{jk}(v_i)}{g_{jk}} \right) \qquad (2.3)$$

b) Graphe dirigé : La même relation peut être utilisée mais doit être multipliée par 2 car un chemin de j vers k est différent du chemin inverse allant de k vers j. De même, g_{jk} considère les chemins dans les deux directions [48].

Dans ce type de centralité, le nombre de liens est moins important que pour le degré même s'il permet de se retrouver sur davantage de chemins géodésiques. Ces positions sont généralement vues comme stratégiques car elles permettent de faire la liaison entre des acteurs ou des groupes d'acteurs qui, autrement, ne seraient que difficilement en contact.

La figure 2.6 indique la centralité d'intermédiarité pour les nœuds du graphe G :

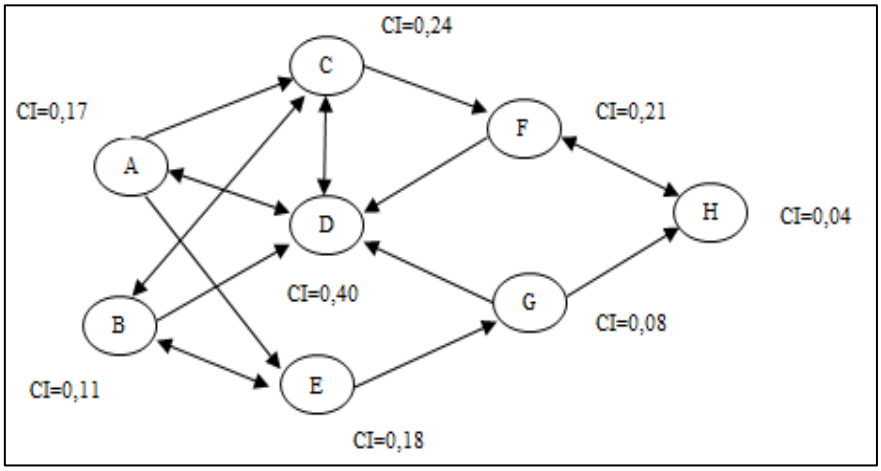

Figure 2.6 : Centralité d'intermédiarité (CI)

Le nœud D possède la plus forte centralité.

Le nœud le plus central est celui par lequel le plus d'acteurs doivent passer quand ils veulent atteindre d'autres acteurs du réseau.

2.3.5 Mesures de centralité issues de la recherche d'information

Les études faite par CHIKHI [54] nous a permis de résumer les algorithmes les plus connus pour le calcul de centralité dans les réseaux sociaux. Ces algorithmes sont tous basés sur l'idée que l'importance d'un document est relative à l'importance des documents auxquels il est connecté.

2.3.5.1 PageRank

L'algorithme PageRank, proposé à la fin des années 90 par les deux informaticiens BRIN et PAGE [56], est à l'origine du moteur de recherche Google. PageRank est l'un des algorithmes d'analyse de liens qui a le plus marqué le domaine de la recherche d'information sur le web.

La mesure PageRank repose sur un concept : un lien émis par une page A vers une page B est assimilé à un vote de A pour B. Plus une page reçoit de votes, plus cette page est considérée comme importante. Plus précisément, le PageRank simplifié d'une page p_i est donné par [57] :

$$PR_s(p_i) = \sum_{p_j \in in(p_i)} \frac{PR_s(p_j)}{d^{out}(p_j)} \quad (2.4)$$

où $in(p_i)$ représente l'ensemble des pages qui pointent vers la page $p(i)$ et $d^{out}(p_j)$ représente le degré sortant de la page p_j.

La figure 2.7 montre les résultats obtenus en appliquant l'algorithme PageRank simplifié. La figure indique que le nœud D est le plus populaire (ou le plus important) car il est pointé par plusieurs nœuds qui sont eux même importants. Nous remarquons aussi que le nœud A possède un PageRank plus important que les nœuds B, E et H bien que ces derniers aient plus de liens entrants que A. Cela s'explique par le fait que A est pointé par un nœud (à savoir D) qui est plus important que les nœuds qui pointent vers B (à savoir C et E), E (à savoir A et B) ou H (à savoir F et G).

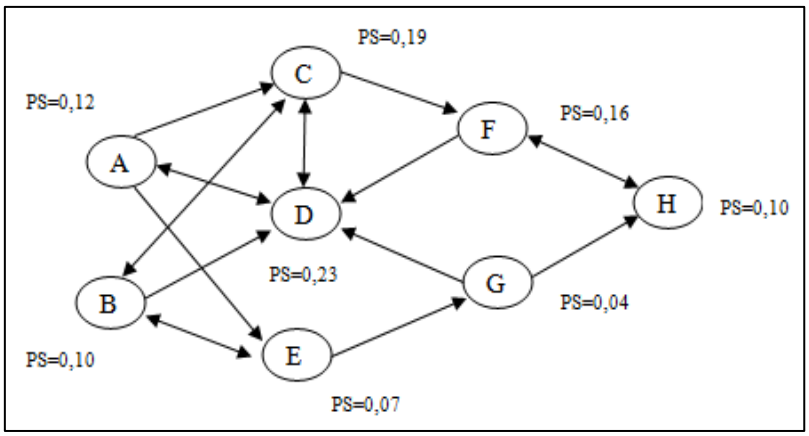

Figure 2.7 : PageRank simplifié (PS)

Le principe de base du PageRank est d'attribuer à chaque page un score proportionnel au nombre de fois que passerait par cette page un utilisateur parcourant le graphe du Web en cliquant aléatoirement sur les liens contenus sur chaque page. Ainsi, une page possèdera un PageRank d'autant plus important que sera grande la somme des PageRank des pages qui pointent vers elle ; une page est importante si elle est pointée par d'autres pages importantes.

2.3.5.2 HITS

Kleinberg a proposé l'algorithme HITS (Hypertext Induced Topic Search) [58 ; 59] pour identifier les documents autorités au moment de la recherche. L'algorithme HITS caractérise chaque page par deux degrés d'importance. Ces deux degrés, que Kleinberg appelle degrés d'autorité et d'hubité, sont respectivement des mesures de centralité par rapport aux liens entrants et aux liens sortants.

L'algorithme HITS considère que le degré d'autorité d'une page est égal à la somme des degrés d'hubité des pages qui la pointent (ou la citent). En d'autres termes, une page est une bonne autorité si elle est pointée par de bons hubs. Plus précisément, le degré d'autorité d'une page p_i est défini par :

$$A(p_i) = \sum_{p_j \in in(p_i)} H(p_j) \quad (2.5)$$

où $in(p_i)$ représente l'ensemble des pages qui pointent vers la page p_i et $H(p_j)$ représente le degré d'hubité de la page p_j.

De manière similaire, HITS considère que le degré d'hubité d'une page est égal à la somme des degrés d'autorité des pages qu'elle pointe. Cela sous-entend qu'une page est un bon hub si elle pointe vers de bonnes autorités. Ainsi, le degré d'hubité d'une page p_i est défini par :

$$H(p_i) = \sum_{p_j \in out(p_i)} A(p_j) \quad (2.6)$$

où $out(p_i)$ représente l'ensemble des pages pointées par la page p_i et $A(p_j)$ représente le degré d'autorité de la page p_j.

La figure 2.8 montre que A et B sont les meilleurs hubs : ils pointent en effet vers plusieurs nœuds ayant un fort degré d'autorité. Notons enfin que les nœuds A

et B ont le même degré d'hubité car ils pointent vers les mêmes nœuds à savoir C, D et E.

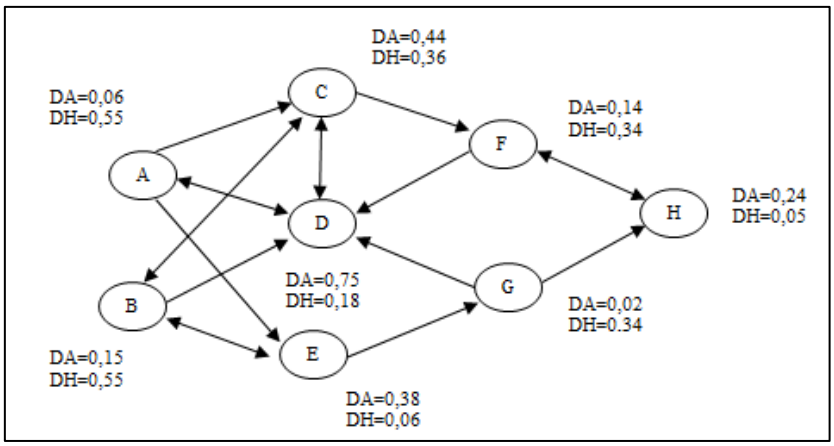

Figure 2.8 : Degrés d'autorité (DA) et d'hubité (DH) calculés par HITS

2.4 La recherche d'information sociale

Les systèmes de recherche d'information classiques se fondent principalement sur une architecture centralisée où l'utilisateur exploite un service de recherche afin de trouver les informations correspondant à son intérêt. Dans ce contexte les utilisateurs ne jouent que le rôle de consommateurs du service dans la recherche, aucune collaboration n'étant nécessaire entre eux [60].

Dans la RIS, le rôle humain devient important. Une personne ne sera plus seulement traitée comme un consommateur mais aussi comme une source de connaissances. Les utilisateurs participeront eux-mêmes au processus de valorisation d'informations dans la recherche.

KIRSCH [61] a défini la recherche d'information sociale par l'intégration des données des réseaux sociaux dans le processus de recherche d'information : « *Social information retrieval systems are distinguished from other types of information retrieval systems by the incorporation of information about social networks and relationships into the information retrieval process* ».

2.4.1 Graphe du contenu social

Les données sociales (les documents, les commentaires, les annotations, ...) ainsi que les interactions mutuelles entre les personnes peuvent être exploitées pour améliorer la recherche d'information. La figure 2.9 montre le graphe du contenu social qui comprend deux entités : les personnes et les données, et quatre types d'interactions (de contenu à contenu, de contenu à personne, de personne à personne et de personne à contenu). Ces interactions définissent le contexte social :

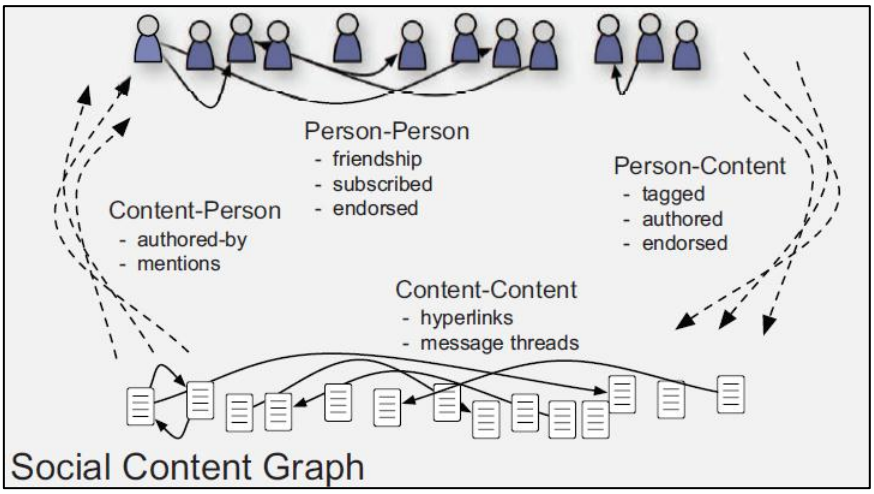

Figure 2.9 : Le graphe du contenu social [62]

- Les interactions de Contenu à Contenu donnent les informations sociales afin de mettre en évidence les documents centraux dans le graphe du Web.
- Les interactions de Contenu à Personne permettant d'identifier les personnes ciblées qui représentent le principal sujet du contenu.
- Les interactions de Personne à Personne aident à identifier les personnes expertes dans un sujet dans le réseau social.
- Les interactions de Personne à Contenu peuvent refléter l'intérêt de l'utilisateur pour le contenu publié.

La topologie du graphe du contenu social peut varier en fonction de la demande sociale, mais doit inclure les deux principaux types d'entités : les acteurs et les données. Par exemple, le réseau d'information sociale de [63] représente les entités sociales qui interagissent au voisinage du document. Dans la figure 2.10, les acteurs représentent les producteurs et les consommateurs d'information

(respectivement les auteurs et les utilisateurs) tandis que les données comprennent les documents et les annotations sociales (les *tags*, les votes, et les avis). Dans le cadre de leurs collaborations et interactions sociales, les acteurs participent à produire de l'information et à enrichir les documents par les annotations sociales.

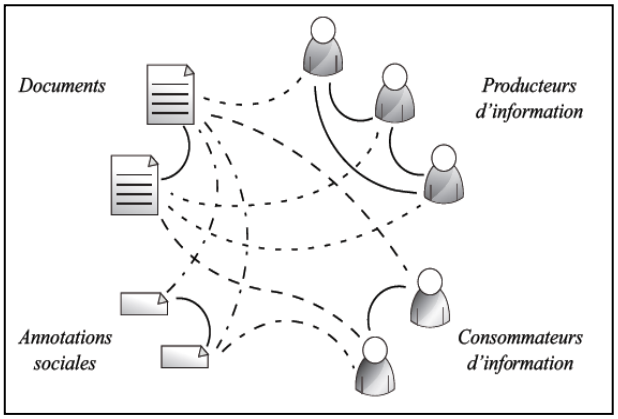

Figure 2.10 : Le réseau d'information sociale [63]

Le réseau d'information sociale peut être formellement représenté par un graphe G = (V,E) où l'ensemble des nouds V = A ∪ U ∪ D ∪ T représente les entités sociales avec A, U, D et T correspondant respectivement aux auteurs, aux utilisateurs, aux documents et aux annotations sociales. L'ensemble des arcs E = (V × V) représente les relations sociales reliant les différents types de nœuds (publier, co-auteur, amitié, citation, annotation, ...).

2.4.2 Système de recherche d'information sociale

La recherche d'information sociale diffère d'autres approches de recherche par l'intégration de la structure du réseau social dans le processus de recherche. Le cycle de recherche comprend trois étapes élémentaires : l'extraction de réseau social, l'analyse de réseau social et le classement par pertinence des documents [64]. Le système de recherche d'information sociale est présenté par la figure 2.11.

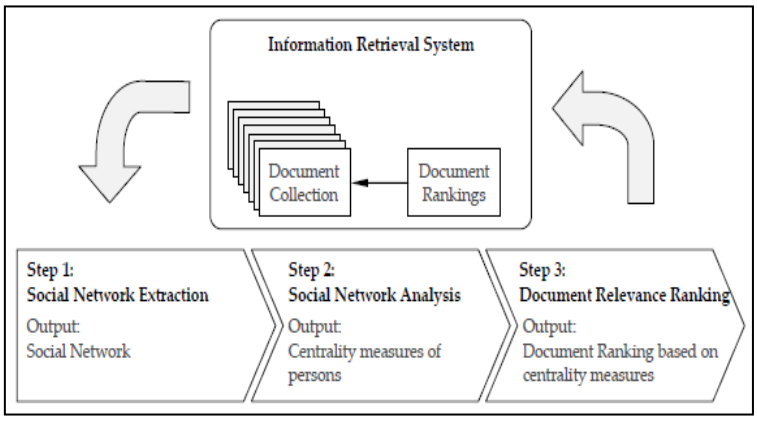

Figure 2.11 : Système de recherche d'information sociale [64]

Etape 1 : Extraction de réseau social. Dans cette étape, la structure du réseau social est extraite de la collection de documents. En dépit des relations sociales explicites, de nouvelles relations sociales sont dérivées du contenu et des métadonnées ; par exemple, les co-citations sont extraites du graphe de citation.

Etape2 : Analyse des réseaux sociaux. Dans cette étape, les méthodes d'analyse de réseaux sociaux sont appliquées sur le réseau social afin d'identifier les principales entités. Un score de pertinence sociale, basé sur les mesures de centralité, est attribué à chaque acteur du réseau.

Etape3 : Classement par pertinence du document. Dans cette étape, un score de pertinence basé sur une requête est combiné avec un score basé sur la pertinence sociale afin de produire un classement final des documents.

2.4.3 Tâches de recherche sociale

De nouvelles approches sociales ont été proposées récemment afin de satisfaire les besoins en information des utilisateurs. Ces approches sont classifiées en cinq grandes catégories selon les tâches de recherche [43] :

a) Recommandation : Au-delà des scores (données explicites) qui sont utilisés dans le filtrage collaboratif, les systèmes de recommandation sociaux s'appuient également sur l'analyse des contenus (données implicites) des interactions entre utilisateurs sur les medias sociaux (commentaires, tags, ...) [65]. Des sites comme

StumbleUpon[3] proposent par exemple des recommandations de pages web à partir des similarités des scores attribués par les utilisateurs, des scores attribués par les amis, et des centres d'intérêts de l'utilisateur et de ses amis sélectionnés dans une liste de près de 500 domaines d'intérêts.

b) Extraction d'information Sociale : L'utilisateur doit en savoir plus sur la structure du réseau social pour mieux comprendre et explorer le graphe social. Ainsi, les approches d'extraction sociale ont étudié les propriétés sociales du graphe du contenu et ont proposé plusieurs modèles d'exploration permettant d'extraire de nouvelles connaissances disponibles sur le réseau social [66 ; 67 ; 68]. Par exemple, les approches de clustering en communautés sont proposées dans ce contexte pour aider les utilisateurs à trouver les personnes ayant des intérêts similaires dans le réseau social [69 ; 70].

c) Recherche d'opinion : blogs, avis et commentaires sont les différents types de contenu persistant sur le Web qui permettent aux gens d'exprimer leur opinion sur les événements, les produits et services [71 ; 72]. Ces informations permettent à d'autres personnes d'apprendre d'une expérience similaire et de prendre ensuite une décision précise. Avant de réserver une chambre d'hôtel, par exemple, les utilisateurs aimeraient consulter les commentaires des clients sur les services de chambre.

d) Recherche de personnes : la recherche dans le réseau social permet aux utilisateurs de trouver d'autres personnes dans le réseau social qui satisfont certains critères ou qui présentent des propriétés sociales particulières (la popularité, l'influence, l'expertise).

e) Recherche sociale : L'approche de recherche sociale assure une tâche de recherche d'information en tenant compte de la structure du réseau social [73]. Par conséquent, les documents sont classés selon leur pertinence ainsi que leur importance dans le réseau social [64 ; 61 ; 43].

[3] www.stumbleupon.com

2.5 Taxonomie de recherche d'information sociale

La recherche d'information sociale a comme objectif d'améliorer le processus de RI en exploitant les informations provenant de réseaux sociaux.

D'après [74], les contributions les plus importantes dans le domaine de la RIS sont classées en trois principales catégories qui sont :

- Recherche Web sociale : dans laquelle l'information sociale est utilisée dans le but d'améliorer le processus de RI classique, par exemple, classement des documents, la réécriture de la requête, le profil des utilisateurs, …
- Recherche sociale : dans laquelle il s'agit de trouver des informations à l'aide de ressources sociales, comme la demande d'aide à des amis, à des bibliothécaires de référence, ou à des inconnus en ligne [75].
- Recommandation sociale : dans laquelle le réseau social de l'utilisateur est utilisé pour faire des recommandations, par exemple, l'aide d'un réseau social de confiance [76].

La figure 2.12 présente cette taxonomie des contributions de la RIS.

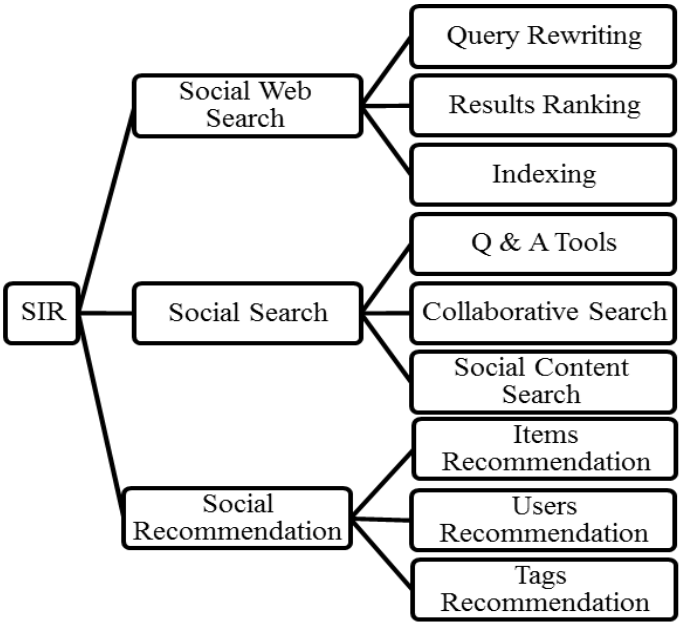

Figure 2.12 : Taxonomie des contributions de la recherche d'information sociale [74]

2.5.1 Recherche Web sociale

Pour améliorer le processus de RI et réduire la quantité de documents non pertinents, il existe principalement trois pistes possibles d'amélioration : i) reformulation de requêtes à l'aide de connaissances supplémentaires, à savoir l'expansion ou l'amélioration de la requête de l'utilisateur, ii) un reclassement des documents récupérés (sur la base du contexte), et iii) l'amélioration du modèle de RI, à savoir la façon dont les documents et les requêtes sont représentés et adaptés à quantifier leurs similitudes.

a) Reformulation de requêtes

La reformulation de requêtes (*Query Reformulation*) est le processus qui consiste à transformer une requête initiale q en une autre requête q'. Cette transformation peut être soit un perfectionnement ou une expansion. Le raffinement de requête réduit la requête de telle sorte que l'information inutile est éliminée, tandis que l'expansion de requête ajoute de nouvelles informations à la requête initiale pour la rendre moins ambigüe.

L'idée est d'exploiter les interactions des utilisateurs avec le système pour construire implicitement et en collaboration une base de données de termes, qui devrait alimenter le processus d'expansion. Cela donnera une source de vocabulaire basé sur l'utilisateur pour l'expansion de la requête [74].

b) Classement des résultats

En RI, le classement des résultats consiste en la définition d'une fonction de classement qui permet de quantifier les similitudes entre les documents et les requêtes. Nous distinguons deux catégories de classement des résultats sociaux qui diffèrent dans la manière dont ils utilisent l'information sociale. La première catégorie utilise l'information sociale par l'ajout d'une pertinence sociale dans le processus de classement, tandis que la seconde utilise l'information sociale pour personnaliser les résultats de recherche.

Les recherches récentes en RI se sont focalisées sur l'utilisation des annotations sociales et/ou relations sociales pour l'amélioration des résultats de RI et le reclassement des documents par l'intégration de ces annotations dans les

différentes méthodes habituelles de calcul de score [77 ; 78]. D'autres travaux proposent de reclasser les ressources ou documents sur le Web en intégrant l'information sociale par combinaison des scores : un score social et un score thématique classique d'un document par rapport à une requête [79], où le score social peut être obtenu à partir des annotations et relations sociales [63].

c) Indexation des documents

Avec l'avènement du Web social où tous les utilisateurs sont contributeurs, les pages Web sont associées à un contexte social qui peut en dire beaucoup sur leur contenu (par exemple, les annotations sociales). Par conséquent, le contexte social est nécessaire pour renforcer le contenu textuel des pages Web. Plusieurs travaux de recherche ont déclaré que l'ajout d'une étiquette sur le contenu d'un document améliore la qualité de la recherche car ils sont de bons résumés de documents [80].

BOUHINI a présenté un modèle de recherche d'information sociale qui intègre le contexte informationnel social des utilisateurs, construit à partir de ses annotations sociales dans la phase d'indexation [49]. Les résultats d'évaluation du modèle de RIS proposé ont montré une amélioration par rapport à la recherche classique.

2.5.2 Recherche sociale

La Recherche Sociale (*Social Search*) est le processus de recherche d'information seulement à l'aide des entités sociales, en tenant compte des interactions ou des contributions des utilisateurs.

Ainsi, la recherche sociale est associée à des plates-formes qui sont définies comme les moteurs de recherche spécifiquement dédiés à la gestion des données sociales telles que Facebook. L'ingrédient principal pour effectuer une recherche sociale sont les interactions de l'utilisateur, y compris: i) le contenu social (par exemple, les commentaires, tweets, ...), et ii) les relations sociales (par exemple, de trouver une personne avec une certaine expertise).

a) Question / Réponse (Q & A) sociale

Malgré le développement des techniques et méthodes pour la recherche sur le Web, de nombreuses requêtes restent sans réponse. DROR affirme que cela est dû

principalement à deux raisons: i) l'intention derrière la requête n'est pas bien exprimée / capturée, et ii) l'absence d'un contenu pertinent [81].

Pour s'attaquer à ces problèmes, les systèmes de Question / Réponse (Q & A) ont vu le jour afin de connecter les utilisateurs entre eux pour s'entraider à répondre aux questions. Aardvark est un des systèmes Q & A [82] permettant de connecter des utilisateurs avec des amis ou des amis-des-amis qui sont en mesure de répondre à leurs questions. Avec Aardvark, les utilisateurs posent une question par messagerie instantanée, e-mail, entrée Web, message texte ou vocal ; Aardvark route alors la question à la personne sur un réseau social étendu de l'utilisateur la plus susceptible d'être en mesure de répondre à cette question. Par rapport à un moteur de recherche Web traditionnel, où le défi consiste à trouver les documents qui sont susceptibles de répondre à la requête de l'utilisateur, le défi dans un moteur de recherche sociale comme Aardvark est de trouver la bonne personne pour satisfaire le besoin d'information de l'utilisateur.

b) Recherche sociale collaborative

L'une des faiblesses des moteurs de recherche disponibles aujourd'hui (par exemple Google, Yahoo !, Bing) est le fait qu'ils soient conçus pour un seul utilisateur qui cherche seul. Ainsi, les utilisateurs ne peuvent pas bénéficier de l'expérience des autres pour une tâche de recherche donnée.

Dans un tel contexte, FILHO [83] a proposé Kolline, une interface de recherche qui vise à faciliter la recherche d'information pour les utilisateurs inexpérimentés en permettant aux utilisateurs plus expérimentés de collaborer ensemble.

Les systèmes de recherche sociale collaborative sont un moyen aidant les utilisateurs à partager leurs expériences et résultats.

c) Recherche du contenu social

Les systèmes de recherche de contenu social permettent d'extraire et de trouver l'information pertinente dans les tags et les commentaires des utilisateurs.

Des plateformes sociales permettent aux utilisateurs de fournir, publier et diffuser de l'information (par exemple, Tweeter) ou de commenter un événement. Dans un tel contexte, une énorme quantité d'informations est créée dans les médias

sociaux, ce qui représente une précieuse source d'information pertinente. Par conséquent, de nombreux utilisateurs utilisent les médias sociaux pour recueillir des informations récentes sur un contenu particulier.

Ainsi, les systèmes de recherches de contenu social sont un moyen d'indexer le contenu explicitement créé par les utilisateurs sur les médias sociaux et de fournir un support de recherche en temps réel [84].

2.5.3 Recommandation sociale

La troisième catégorie des contributions de la RIS considère le champ de filtrage et de recommandation (filtrage basé sur le contenu par exemple, filtrage collaboratif, les systèmes de recommandation).

Les systèmes de recommandation sociaux s'appuient sur différents types de réseaux sociaux extraits des medias sociaux ou d'outils collaboratifs tels que les réseaux sociaux en ligne, les sites de social bookmarking, les wikis, les blogs, ... [85]. Les réseaux sociaux exploités peuvent être déduits des interactions directes entre utilisateurs (score de confiance, amitié, tags entre utilisateurs, organigramme d'entreprise, ...) ou des interactions entre utilisateurs et ressources (co-auteur d'un même article, commentaires ou tags sur une même page wiki, ...) tels que proposé par [86].

Au-delà des scores (données explicites) qui sont utilisés dans le filtrage collaboratif, les systèmes de recommandation sociaux s'appuient également sur l'analyse des contenus (données implicites) des interactions entre utilisateurs sur les medias sociaux (commentaires, tags, ...) [65].

GUY [86] intègre (voir figure 2.13) les contenus des tags en plus des relations entre personnes et items, pour prédire les scores (poids) entre utilisateurs et items en fonction de leur réseau social.

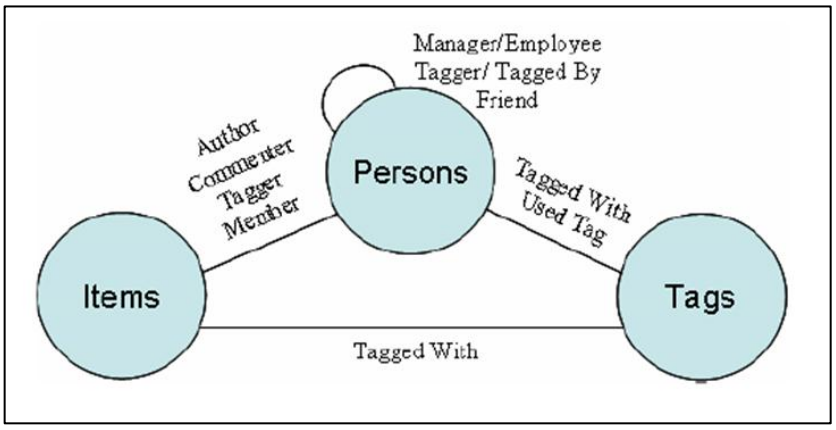

Figure 2.13 : Exemple d'interactions directes et indirectes pour la construction d'un réseau social [87 ; 86]

La recommandation sociale (*Social Recommandation*) est un ensemble de techniques qui tentent de suggérer : i) du contenu (par exemple, des films, de la musique, des livres, des nouvelles, des pages Web), ii) des entités sociales (par exemple, les personnes, les événements, les groupes), ou iii) des sujets d'intérêt (par exemple, le sport, la culture, la cuisine, …) qui sont susceptibles de présenter un intérêt pour l'utilisateur grâce à l'utilisation de l'information sociale.

BOUADJENEK [74] classe les systèmes de recommandation sociaux en trois types :

a) Recommandation d'articles

Les méthodes de recommandation d'articles classiques sont basées sur l'hypothèse que les utilisateurs sont des entités indépendantes et ne supposent aucune structure, y compris le réseau social qui entoure les utilisateurs. Cela ne reflète pas les comportements réels des utilisateurs car ils demandent parfois à leurs amis des recommandations avant d'agir, par exemple pour l'achat d'un produit.

b) Recommandation d'utilisateurs

Les plates-formes de réseaux sociaux ont adopté la stratégie de suggérer des amis (ou groupe d'amis) pour augmenter la connectivité entre leurs utilisateurs. WANG [88] a proposé une approche pour connecter les utilisateurs avec des goûts similaires en mesurant leurs similitudes sur la base des tags qu'ils partagent dans un réseau utilisant les tags.

c) Recommandation de sujets

Récemment, la recommandation de sujet et d'étiquette a attiré une attention considérable. La recommandation des tags permet également aux utilisateurs de choisir les bons tags.

Referral Web [89] modélise un réseau social en analysant les sources de communication (e-mail, …) pour obtenir un modèle du réseau. Une fois construit, le réseau social peut être parcouru et des informations sur des personnes parlant d'un sujet en particulier peuvent être extraites, par exemple, la liste de documents en rapport avec Michel Smith [90].

2.6 Utilisation des Bookmarks sociaux dans la recherche d'information

Les bookmarks sociaux sont une des applications les plus caractéristiques du Web 2.0 et en sont les précurseurs. Apparus en 2003, ces services offrent la possibilité de partager les bookmarks c'est-à-dire les favoris (titre, adresse et description d'une page ou site). Ainsi, après enregistrement, le plus souvent gratuit, ces favoris mis en ligne peuvent être accessibles aux internautes du monde entier.

Il existe plusieurs dizaines de services de bookmarks sociaux. Nous pouvons citer *Del.icio.us*, qui appartient à *Yahoo! e*t qui est l'un des plus connus et des plus utilisés. *Connotea*, édité par la célèbre revue *Nature,* est destiné à un public scientifique. Lors de la mise en favoris, ce service extrait automatiquement les références bibliographiques lorsqu'elles sont issues de sites tels que *Nature, Science ou PubMed*. Plus ancien, *CiteUlike* est basé sur le même principe mais est "compatible" avec davantage de revues scientifiques. *Snipitron* est lui dédié aux chercheurs, étudiants et professionnels.

Le Web social contient une grande quantité de données générées par les utilisateurs. Ces données sont des statuts, des commentaires, des tags, ... Parmi ces données, dans un réseau social, les tags sont les plus utilisés pour la recommandation. Un tag (ou étiquette) est un mot-clé ou terme associé ou assigné à de l'information (par exemple une image, un article, ou un clip vidéo). Les tags sont habituellement choisis de façon informelle et personnelle par le créateur ou le consommateur de la ressource. De nombreux réseaux sociaux permettent aujourd'hui

aux utilisateurs d'ajouter des tags aux objets pour qu'ils puissent les retrouver facilement plus tard.

Le terme tag est lié à la notion de folksonomie (*folksonomy* en anglais) qui désigne un système de classification collaborative basé sur les tags.

2.6.1 Opération d'étiquetage

À l'utilisation de tags est liée la pratique d'étiquetage ou de tagging, c'est-à-dire l'association par un utilisateur d'un tag à une ressource donnée. Cette relation qui forme ainsi une relation tripartite [91] peut se représenter par *Tagging* (*Utilisateur, Ressource, Tag*) telle que :

- *Utilisateur* correspond à l'utilisateur qui effectue l'action ;
- *Ressource* correspond à la ressource annotée (billet de blog, page Web ...) ;
- *Tag* correspond au tag utilisé ;
- *Tagging* correspond à l'action liant ces trois éléments.

Étant donné que plusieurs tags peuvent être associés par un même utilisateur à une même ressource, et qu'un même tag peut être associé à une même ressource par différents utilisateurs, les actions de tagging ne sont en général pas isolées (voir figure 1.14). On utilise donc l'appellation de social tagging ou de métadonnée sociale.

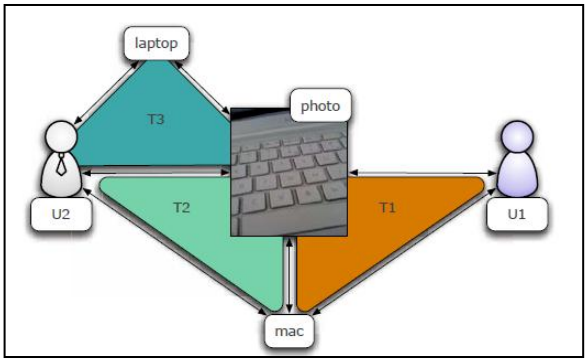

Figure 2.14 : Actions de tagging combinées autour d'une même photo [44].

La figure 2.14 présente trois actions de tagging (*T*1, *T*2, *T*3) associées à une même ressource (*photo*) via deux utilisateurs (*U*1, *U*2) et deux tags distincts (*mac*, *laptop*) de la manière suivante :

- *T*1(*U*1*, photo, mac*)
- *T*2(*U2, photo, mac*)
- *T*3(*U2, photo, laptop*)

2.6.2 Recommandations

Le réseau social peut être aussi une source permettant l'élaboration de recommandations. La collaboration des services de social bookmarking et le rôle des tags pour décrire à la fois le contenu des signets et les préférences de l'utilisateur ont fait des données une ressource naturelle pour la génération de recommandations pertinentes [92].

Les utilisateurs sont les acteurs principaux du système de folksonomie et contribuent au contenu par l'ajout de ressources et l'affectation de tags. Cependant, il s'avère que le choix des tags et des ressources partagées par un utilisateur d'une folksonomie varie selon plusieurs critères : le genre, l'âge ou encore la profession de celui qui partage l'information. Ainsi, les folksonomies doivent tenir compte de telles informations lors de la recommandation de tags ou de ressources [93].

Ainsi, un système de recommandation offre à l'utilisateur une liste de tags ou de ressources recommandés qui lui permet de trouver plus facilement ses tags et ressources préférés dans la folksonomie [94].

Dans un souci d'améliorer les recommandations dans les folksonomies, plusieurs travaux ont été proposés dans la littérature. Dans [95], les auteurs utilisent les tags d'un utilisateur afin de lui recommander des utilisateurs ayant partagé des tags et des ressources similaires. Dans [96], les auteurs se basent à la fois sur l'historique de tagging (tags et ressources) des utilisateurs et sur leurs contacts sociaux. Dans [97], JÂSCHKE a proposé des recommandations de tags dans les folksonomies basées sur les tags les plus utilisés. LIPCZAK [98] a proposé un système de recommandation de tags en trois étapes : les tags de base sont extraits du titre de la ressource, puis une extension de l'ensemble des recommandations est faite par l'ajout des tags proposés par un lexique sur la base de cooccurrences de tags dans les contenus des ressources ; ensuite, les tags sont filtrés selon le besoin de l'utilisateur.

2.7 Conclusion

Nous avons présenté dans ce chapitre un aperçu sur la recherche d'information sociale. En outre, nous avons donné les principales mesures de centralité qui évaluent l'importance des acteurs sociaux. Enfin, nous avons discuté les tâches de recherche d'information sociale en accordant une attention particulière à la recherche Web sociale. Cette tâche exploite la structure de réseau social afin d'améliorer le processus de classement du système de recherche d'information.

Dans le chapitre suivant, nous présentons un nouveau modèle social des utilisateurs du réseau de bookmarking où les relations sont extraites à partir des activités de co-marquage.

CHAPITRE 3
APPROCHE PROPOSEE POUR LA RECHERCHE D'INFORMATION SOCIALE

3.1 Introduction

Les documents sont consommés par des entités sociales et leur importance peut être estimée à partir du contexte d'utilisation. A cet effet, l'introduction de la dimension sociale dans le processus de RI est un plus qui peut améliorer la qualité des documents retournés.

Les informations sociales (relations sociales, annotations, clics, profils, ...) peuvent être exploitées au sein même du modèle de RI (modèle de document et de requête, fonction de pondération / de correspondance), ou en aval de ce modèle (reclassement de la liste des résultats).

Notre contribution s'inscrit dans le domaine de recherche Web sociale (voir figure 3.1). Plus précisément, nous proposons une approche utilisant l'analyse des réseaux sociaux pour le reclassement des résultats retournés par un moteur de RI classique (i.e. basé sur le contenu uniquement).

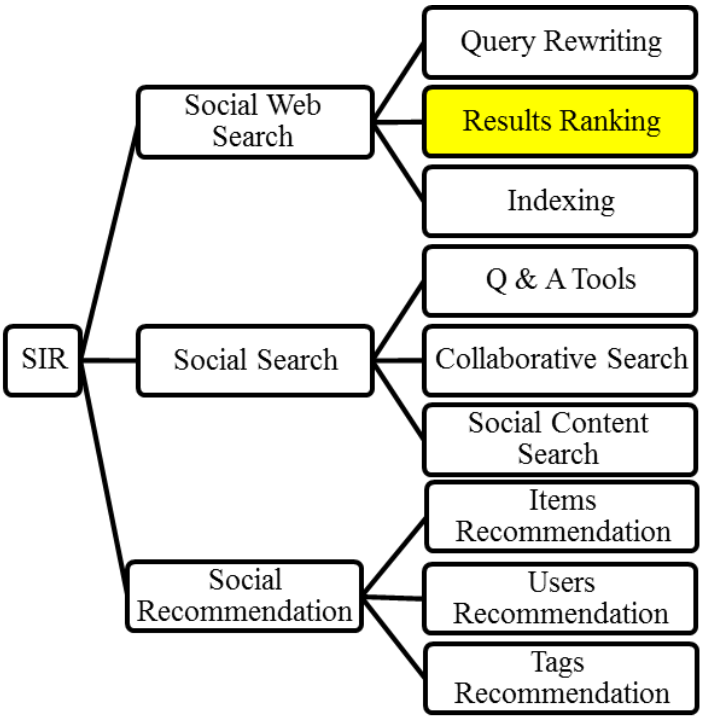

Figure 3.1 : Classement de notre approche

3.2 Architecture générale de l'approche proposée

La figure 3.2 décrit l'architecture générale de l'approche proposée qui s'articule autour de trois modules. Il s'agit dans un premier temps d'extraire, à partir de la collection de données, le contenu social, puis de calculer le score de pertinence social de chaque document. Dans un deuxième temps, d'envoyer une requête au module de recherche qui rend la liste initiale des résultats. Enfin le module de reclassement prend en charge la pertinence issus du contexte de la requête et du contexte du réseau social afin d'améliorer la précision des résultats de recherche. Nous décrivons dans ce qui suit chacun de ces modules en donnant ses différents composants et son principe de fonctionnement.

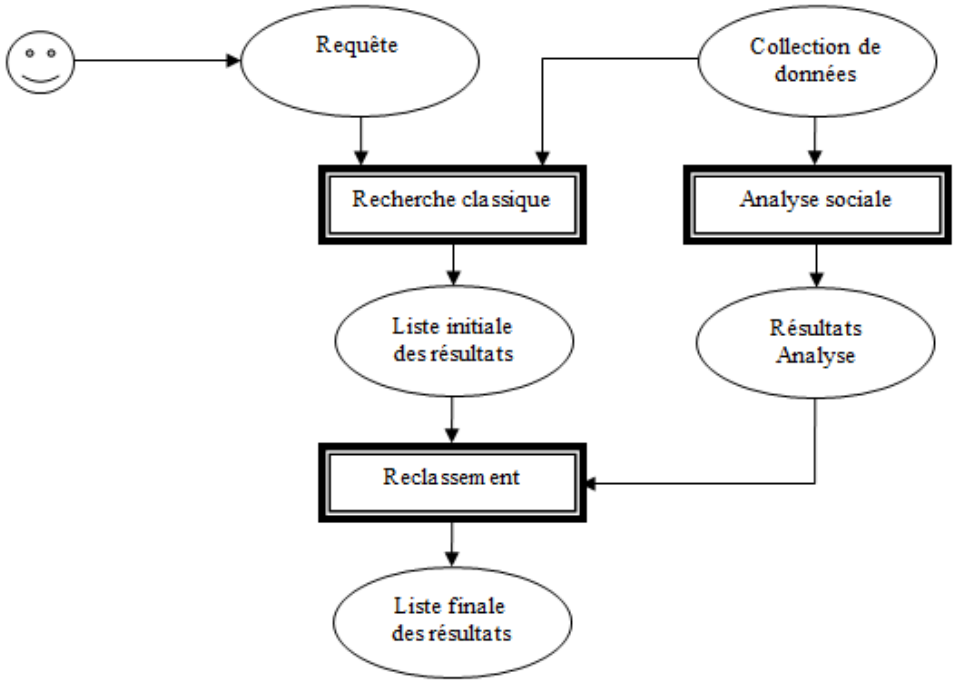

Figure 3.2 : Architecture générale

3.2.1 Recherche classique

Ce module transmet la requête utilisateur au moteur de recherche et récupère les résultats retournés (voir figure 3.3). La liste de résultats récupérée contient le classement de chaque document ainsi que son score.

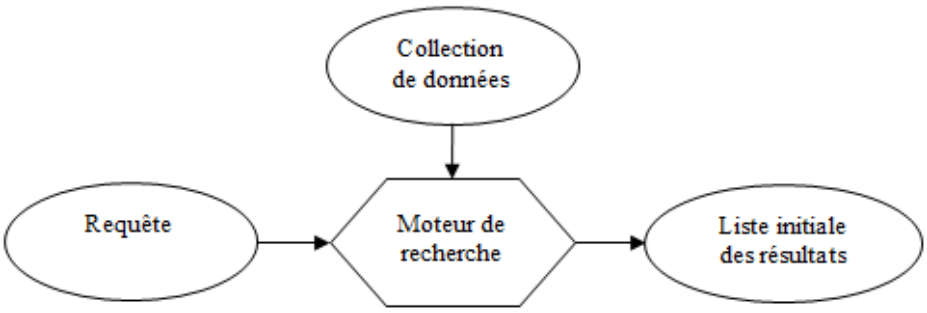

Figure 3.3 : Recherche classique

Le moteur de recherche permet de classer les résultats selon un score de pertinence correspondant à la similarité entre la requête et le contenu textuel des documents.

3.2.2 Analyse sociale

Dans la recherche d'information sociale, les acteurs sont représentés au moyen d'un réseau social et la pertinence sociale d'un document est calculée en appliquant une mesure de centralité issue de l'analyse des réseaux sociaux [99].

Le module d'analyse sociale procède en deux étapes (voir figure 3.4) :

(i) *Construction du réseau d'information sociale.* Dans cette étape, la structure du réseau d'information social est extraite de la collection de données. Ce processus récupère les métadonnées de la base documentaire afin d'extraire les acteurs et les relations du réseau social.

(ii) *Analyse du réseau d'information sociale*. Cette étape permet de calculer le score de pertinence sociale de chaque document à partir du réseau d'information sociale.

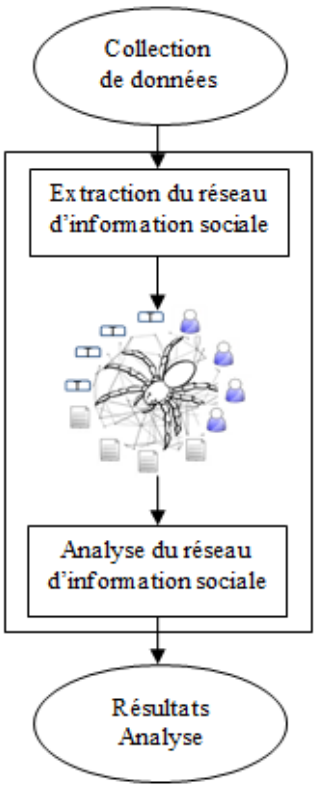

Figure 3.4 : Analyse sociale

3.2.3 Reclassement

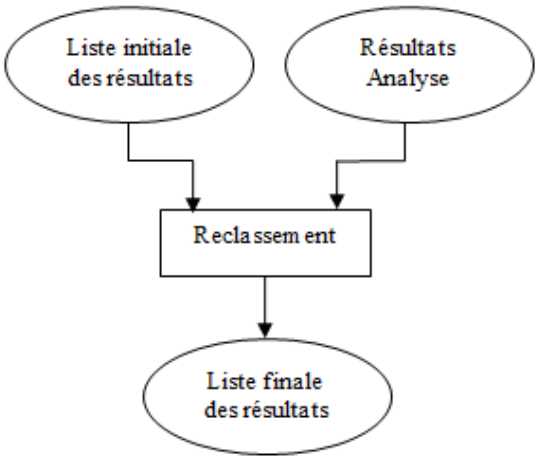

Figure 3.5 : Reclassement

A cette étape, chaque document est décrit par deux valeurs, la première générée par le moteur de recherche et la seconde par l'analyse du réseau d'information social. Le module de reclassement combine les deux scores et procède à l'ordonnancement des résultats suivant le nouveau score.

3.3 L'analyse sociale

Les collections de données sur lesquelles nous travaillons nous permettent d'extraire le contenu social, de l'analyser, puis de calculer le score de pertinence social de chaque document.

3.3.1 Construction du réseau d'information sociale

Formellement, et en se basant sur la représentation proposée par TAMINE et BEN JABEUR [63], notre réseau d'information sociale est représenté par un graphe G = (V,E) où l'ensemble des nœuds V = U ∪ D ∪ T représente les entités sociales avec U, D et T correspondant respectivement aux annotateurs/utilisateurs, aux documents et aux annotations sociales. L'ensemble des arcs E ⊆ (V×V) représente les relations sociales reliant les différents types de nœuds.

Notre réseau d'information sociale représente le contexte d'utilisation sociale des documents et l'interaction entre les utilisateurs.

La figure 3.6 représente notre réseau d'information sociale qui intègre les trois types de nœuds.

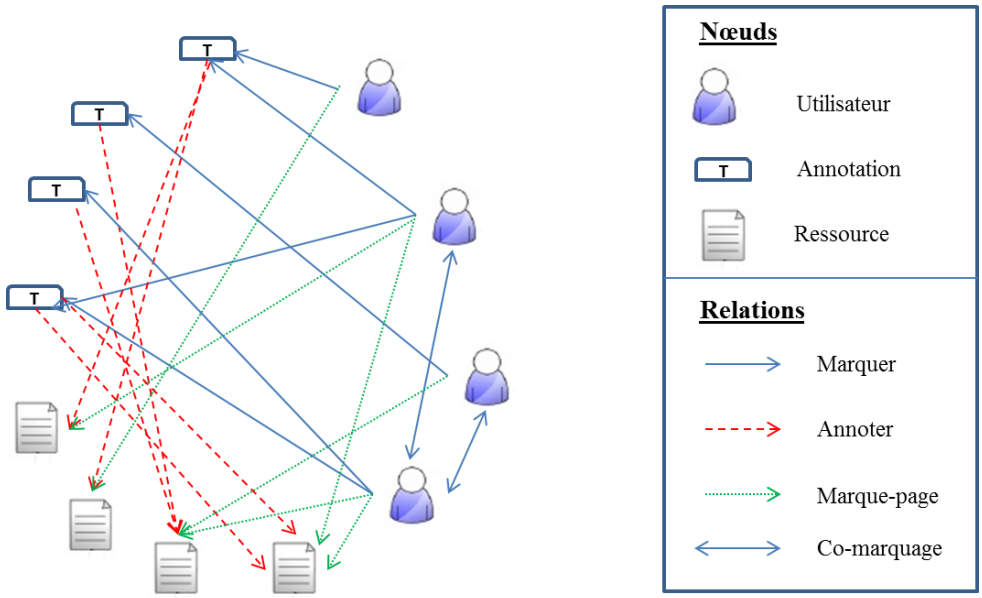

Figure 3.6 : Le réseau d'information sociale

Nous identifions les relations sociales suivantes qui concernent les documents, les *tags* et les utilisateurs :

- Marquer : relie un utilisateur $u_i \in U$ et un *tag* $t_j \in T$ utilisé au moins une fois pour marquer un document.
- Annoter : relie un *tag* $t_j \in T$ avec un document $d_i \in D$ assigné au moins une fois pour décrire son contenu.
- Marque-page : en attribuant un *tag* à un document, l'utilisateur $u_i \in U$ et le document $d_j \in D$ sont associés avec une relation sociale de marque-page.
- Co-marquage : une relation entre deux utilisateurs $u_i \in U$, $u_j \in U$ ayant marqué au moins un document en commun.

3.3.2 Analyse du réseau d'information sociale

Afin d'évaluer la pertinence sociale des documents, nous proposons une approche d'analyse du réseau d'information sociale qui procède en deux étapes (voir figure 3.7) : dans la première étape un réseau social des utilisateurs est extrait puis analysé à l'aide d'une mesure de centralité ; les résultats de cette étape sont ensuite utilisés dans la deuxième étape afin d'assigner un score de pertinence sociale à chaque document.

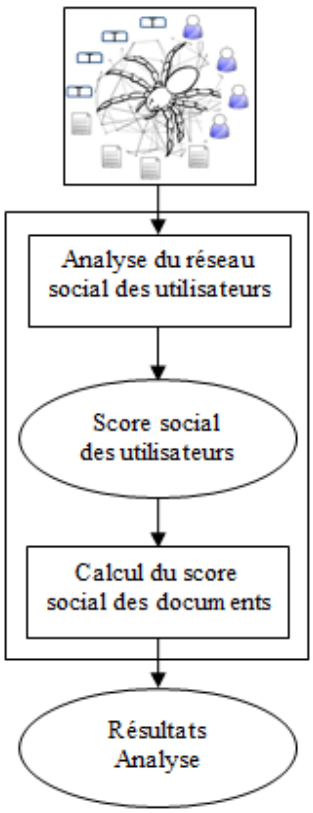

Figure 3.7 : Analyse du réseau d'information sociale

3.3.2.1 Analyse du réseau social des utilisateurs

Nous proposons deux méthodes différentes pour l'analyse du réseau social des utilisateurs.

La première méthode s'intéresse uniquement aux utilisateurs et aux documents tandis que la deuxième exploite toutes les informations du réseau à savoir les utilisateurs, les documents et les tags.

a) Méthode 1 **:** Analyse du réseau social des utilisateurs sans prise en compte des tags

La relation de co-marquage est représentée par un lien ; cette relation connecte deux utilisateurs ayant collaboré pour marquer un document (la figure 3.8 illustre un exemple d'extraction du réseau social). Les utilisateurs ont des relations indirectes à travers les ressources qu'ils marquent.

Le co-marquage exprime l'intérêt commun entre les utilisateurs.

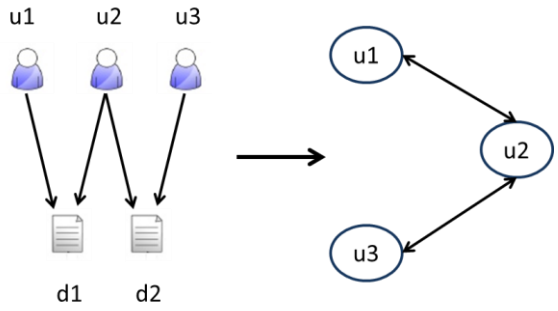

Figure 3.8 : Extraction du réseau social sans prise en compte des tags

Dans cette méthode, nous proposons trois façons pour la pondération des liens du réseau social :

1. Réseau non pondéré

C'est un modèle simple qui est souvent utilisé dans l'analyse des réseaux sociaux.

Dans ce réseau, le poids de chaque lien est égal à 1.

Le réseau social des utilisateurs est ainsi représenté par un graphe G = (U,E) où l'ensemble des nœuds U représente les utilisateurs et l'ensemble des arêtes E ⊆ U×U représente les relations sociales entre eux (i.e. les relations de co-marquage).

Un réseau non orienté peut être représenté par un réseau orienté par liaison symétrique (exemple : voir figure 3.9), c'est-à-dire chaque arête dans le réseau non orienté est remplacée par deux arcs orientés.

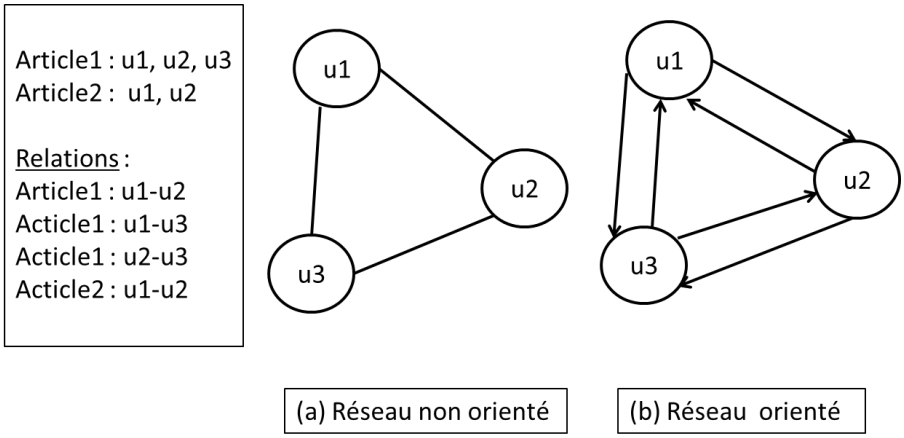

Figure 3.9 : Réseau non pondéré (sans prise en compte des tags)

2. Réseau non-orienté pondéré

Dans le cas d'un réseau pondéré, le réseau social des utilisateurs est représenté par un graphe G = (U,E,P) où l'ensemble des nœuds U représente les utilisateurs, l'ensemble des arêtes E ⊆ U×U représente les relations de co-marquage entre eux et l'ensemble P représente les poids associés à chaque relation d'une paire d'utilisateurs. Ces poids sont normalisés pour avoir des valeurs comprises entre 0 et 1.

Afin de quantifier la collaboration (l'intérêt commun) entre les co-utilisateurs, nous proposons de tenir compte de la totalité des collaborations.

Nous présentons ici les relations de co-marquage par des arêtes.

Afin de prendre en compte les poids des arêtes, nous proposons la formule suivante **(indice de Jaccard[4])** :

$$\mathrm{Co}(i,j) = \frac{w(i,j)}{w(i)+w(j)-w(i,j)} \qquad (3.1)$$

avec w(i,j) le nombre de documents co-marqués par les utilisateurs u_i et u_j ; w(i) et w(j) représentent respectivement le nombre de collaborations de l'utilisateur u_i et de l'utilisateur u_j dans l'opération de co-marquage des documents.

Pour ce modèle, un exemple de réseau social avec les poids associés aux relations est présenté par la figure 3.10 (a).

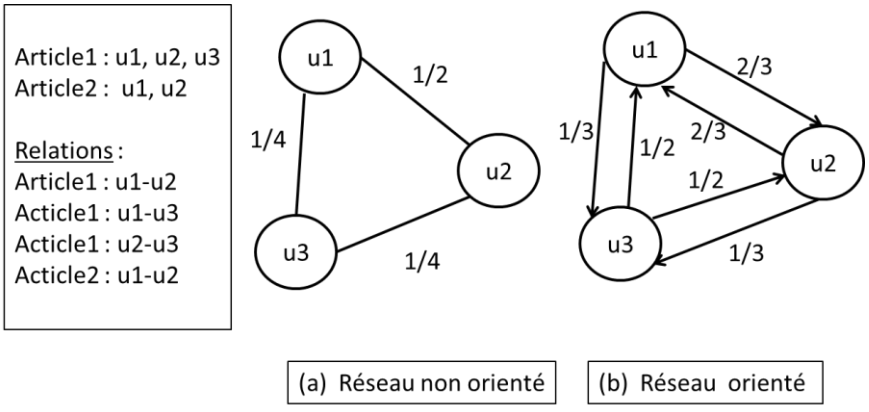

Figure 3.10 : Réseaux pondérés (sans prise en compte des tags)

[4] http://en.wikipedia.org/wiki/Jaccard_index

3. Réseau orienté pondéré

Dans ce cas, la relation de co-marquage est représentée par un arc orienté.

Nous proposons d'assigner des poids aux relations de co-marquage comme suit :

$$\mathrm{Co}(\mathrm{i},\mathrm{j}) = \frac{\mathrm{w}(\mathrm{i},\mathrm{j})}{\mathrm{w}(\mathrm{i})} \qquad (3.2)$$

avec w(i,j) le nombre de documents co-marqués par les utilisateurs u_i et u_j ; w(i) représente le nombre de collaborations de l'utilisateur u_i dans l'opération de co-marquage des documents.

Cette normalisation garantit que la somme des poids des liens sortants d'un nœud est égale à 1.

Pour ce modèle nous présentons un exemple de réseau social avec les poids associés aux relations sur la figure 3.10 (b).

Pour la construction du réseau social et le calcul des poids des liens, nous calculons les paramètres w(i) et w(i,j) par l'algorithme suivant :

Algorithme calcul1

Données :

U = ensemble d'utilisateurs ;

D = ensemble de documents ;

Résultats :

w(i) = Nombre de collaborations de l'utilisateur u_i dans l'opération de co-marquage ;

w(i,j) = Nombre de documents co-marqués par l'utilisateur u_i et l'utilisateur u_j ;

```
Début
    Pour chaque utilisateur u_i ϵ U faire
        w(i)←0 ;
        Pour chaque utilisateur u_j (u_j ≠ u_i) ϵ U faire
                w(i,j)←0 ;
                Pour chaque document d_k ϵ D marqué par u_i faire
                        Si d_k est marqué par u_j
                                w(i)← w(i)+1 ; w(i,j)←w(i,j)+1 ;
                        FSi
                FFaire
        FFaire
    FFaire
Fin
```

b) Méthode 2 : Analyse du réseau social des utilisateurs avec prise en compte des tags

Dans cette méthode, nous considérons que plus les utilisateurs co-marquent avec des tags communs plus leur relation est de poids élevé et par conséquent, ils ont une forte liaison (intérêt commun).

Dans ce réseau, la relation sociale de co-marquage intègre les documents, les tags et les utilisateurs ; cette relation connecte deux utilisateurs ayant utilisé au moins un tag en commun pour marquer au moins un document en commun (la figure 3.11 illustre un exemple d'extraction du réseau social).

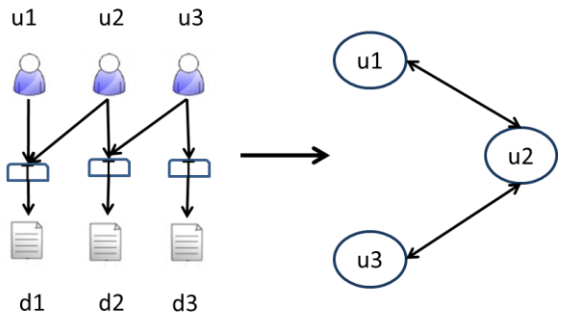

Figure 3.11 : Extraction du réseau social avec prise en compte des tags

Dans cette méthode, nous proposons trois façons pour la pondération des liens du réseau social :

1. Réseau non pondéré

Pour ce modèle de réseau, un exemple est illustré dans la figure 3.12.

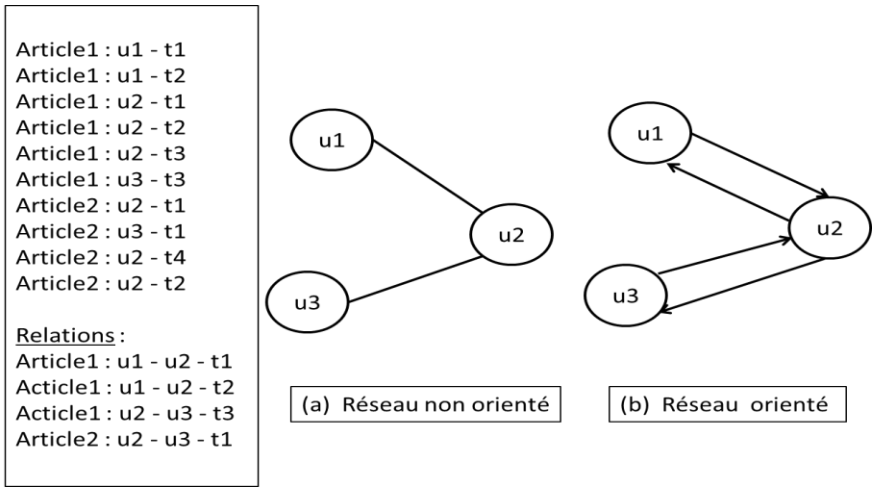

Figure 3.12 : Réseau non pondéré (avec prise en compte des tags)

2. Réseau non-orienté pondéré

Nous proposons d'assigner des poids aux relations de co-marquage comme suit **(indice de Jaccard)** :

$$\mathrm{Co}(i,j) = \frac{w_t(i,j)}{w_t(i)+w_t(j)-w_t(i,j)} \qquad (3.3)$$

avec $w_t(i,j)$ le nombre de tags en commun utilisés par les utilisateurs u_i et u_j pour marquer au moins un document en commun ; $w_t(i)$ et $w_t(j)$ représentent respectivement le nombre de collaborations de l'utilisateur u_i et l'utilisateur u_j dans l'opération de co-marquage des documents.

Pour ce modèle, un exemple de réseau social avec les poids associés aux relations est présenté par la figure 3.13 (a).

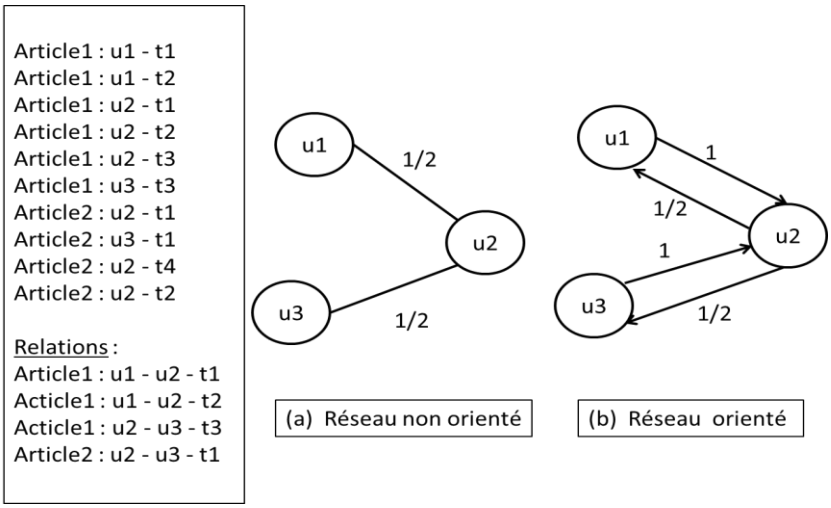

Figure 3.13 : Réseaux pondérés (avec prise en compte des tags)

3. Réseau orienté pondéré

Nous proposons d'assigner des poids aux relations de co-marquage comme suit :

$$\mathrm{Co}(i,j) = \frac{w_t(i,j)}{w_t(i)} \qquad (3.4)$$

avec $w_t(i,j)$ le nombre de tags en communs utilisés par les utilisateurs u_i et u_j pour marquer au moins un document en commun ; $w_t(i)$ représente le nombre de collaborations de l'utilisateur u_i dans l'opération de co-marquage des documents.

Pour ce modèle nous présentons un exemple de réseau social avec les poids associés aux relations sur la figure 3.13 (b).

Cet algorithme explique le calcul des paramètres $w_t(i)$ et $w_t(i,j)$ pour le calcul des poids des différentes relations de co-marquage.

Algorithme calcul2

Données :

U = ensemble d'utilisateurs;

T = ensemble des tags;

D = ensemble de documents.

Résultats :

$w_t(i)$ = Nombre de tags utilisés par l'utilisateur u_i dans l'opération de co-marquage ;

$w_t(i,j)$ = Nombre de tags utilisés par l'utilisateur u_i et l'utilisateur u_j pour co-marquer un document.

```
Début
    Pour chaque utilisateur u_i ϵ U faire
        w_t(i)←0 ;
        Pour chaque utilisateur u_j (u_j ≠ u_i) ϵ U faire
                w_t(i,j)←0 ;
                Pour chaque document d_k ϵ D marqué par u_i faire
                        Si d_k est marqué par u_j
                                Pour chaque tag t_z ϵ T utilisé par u_i faire
                                        Si t_z est utilisé par u_j
                                                w_t(i)← w_t(i)+1 ; w_t(i,j)←w_t(i,j)+1 ;
                                        FSi
                                FFaire
                        FSi
                FFaire
        FFaire
    FFaire
Fin
```

3.3.2.2 Calcul du score de pertinence sociale des documents

La pertinence sociale d'un document est estimée à travers l'importance de ses utilisateurs. Notre objectif est de sélectionner la mesure d'importance sociale (parmi les mesures *Betweenness*, *Closeness* et *PageRank)* la plus appropriée pour calculer, pour chaque utilisateur u_i un score d'importance sociale $C_G(u_i)$.

Ces mesures sont appliquées seulement sur le sous-graphe d'utilisateurs : G = (U,E) avec E ⊆ (U×U).

Un score est calculé pour chaque acteur du réseau social. Cependant, un document peut être lié à de nombreux acteurs. Le score de pertinence sociale attribué à chaque document est calculé par la somme des scores sociaux de ses utilisateurs :

$$\mathrm{Imp_G}(d) = \sum_{i=1}^{m} C_G\,(u_i) \qquad (3.5)$$

avec m : le nombre d'utilisateurs qui ont marqué le document d.

3.4 Le reclassement des résultats

Dans cette étape, un score de pertinence est assigné à chaque document par une combinaison de son score classique et social.

En se basant sur le nouveau score, le système procède à l'ordonnancement des résultats (classement final des documents).

Le rôle de la fonction de classement est d'ordonner les documents avant de les renvoyés à l'utilisateur. En effet, l'utilisateur se contente généralement d'examiner les premiers documents retournés. Par conséquent, si les documents attendus ne sont pas présents dans cette tranche de résultats, l'utilisateur considérera le modèle comme mauvais par rapport à son besoin en information, et les résultats qu'il retourne seront donc considérés comme non pertinents.

3.5 Mise en œuvre de l'approche proposée

3.5.1 Calcul de la pertinence classique

Pour le calcul de la pertinence classique, nous utilisons l'API[5] LUCENE [100]. En fait, lors d'une recherche dans un index de LUCENE, une série ordonnée de résultats sera retournée. Par défaut, ces résultats seront ordonnés selon un score compris entre 0 et 1.

Ce score est calculé pour chaque document de l'index en tenant compte de divers facteurs :

- la fréquence du terme recherché dans le document ;
- la fréquence inverse du terme ;
- les valeurs de normalisation et de correction basées sur le nombre de termes compris dans un champ, l'importance accordée à un champ en particulier, ...

[5] API : Application Programming Interface.

3.5.2 Calcul de la pertinence sociale

Pour le calcul de la pertinence sociale, nous choisissons l'outil d'analyse des réseaux sociaux VISONE (VIsual SOcial NEtworks) [101] qui propose de visualiser directement les graphes décrits par les données, et offre un grand choix pour le calcul de mesures de centralité sur un réseau donné.

Avec le logiciel VISONE, l'analyse des réseaux sociaux se fait en trois temps distincts : la récupération des données, leur analyse et leur visualisation.

3.5.3 Calcul de la pertinence globale

Afin d'améliorer le processus de RI, nous proposons de combiner deux pertinence classique et social

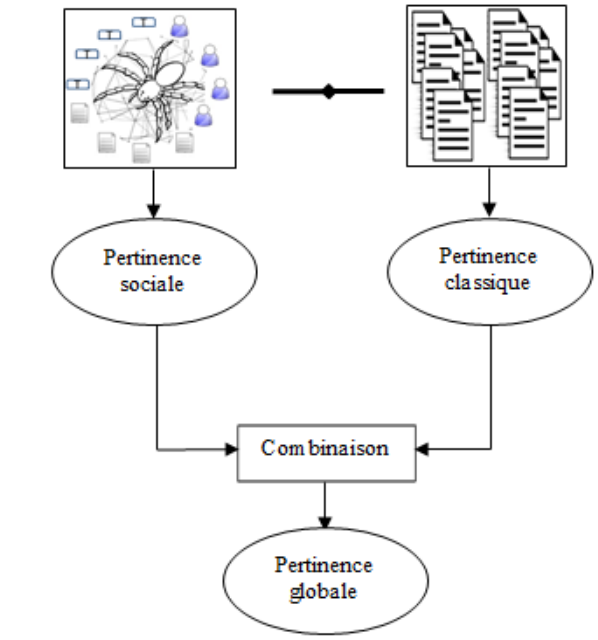

Figure 3.14 : Pertinence globale du document

La pertinence globale $R(q,d,G)$ considère la requête q, le document d ainsi que les interactions des acteurs du réseau social G. En effet, le score de la pertinence globale du document est une combinaison de deux scores de pertinence, le score social $Imp_G(d)$ qui représente l'importance sociale du document

d dans le réseau social G et le score classique RSV[6](q,d), qui représente le degré de similarité entre la requête q et le document d.

Le score de la pertinence globale est calculé comme suit [61] :

$$\mathrm{R}(\mathrm{q},\mathrm{d},\mathrm{G}) = \mathrm{RSV}(\mathrm{q},\mathrm{d}) \times \mathrm{Imp}_{\mathrm{G}}(\mathrm{d}) \tag{3.6}$$

3.6 Conclusion

Notre contribution s'inscrit dans le domaine de la recherche d'information dite contextuelle car elle prend en compte la notion de contexte via le réseau d'utilisateurs pour reclasser les résultats retournés initialement par un système de RI classique.

Nous avons proposé une approche qui intègre l'analyse des réseaux sociaux dans le processus de recherche d'information. En particulier, nous avons présenté le contenu social par un réseau d'information sociale dont les utilisateurs sont les principales entités et les relations sont extraites à partir des liens de co-marquage. Dans cette approche, la pertinence d'un document est estimée par combinaison de la pertinence classique et de la pertinence sociale, qui est à son tour dérivée de l'importance sociale des utilisateurs associés.

Dans le chapitre suivant, nous allons présenter une série d'expérimentations sur un ensemble d'articles de Wikipédia.

[6] RSV : Relevance Status Value.

CHAPITRE 4
EVALUATION EXPERIMENTALE DE L'APPROCHE PROPOSEE

4.1 Introduction

Nous menons une série d'expérimentations sur une collection de 16608 articles Wiki afin d'évaluer notre approche de recherche d'information sociale.

Les principaux objectifs de cette évaluation sont de :

- Comparer les différentes mesures d'importance sociale afin de déterminer celle qui exprime le mieux l'importance des documents ;
- Mesurer l'impact de la pondération des relations sociales sur l'estimation de la pertinence sociale des documents ;
- Déterminer la meilleure méthode d'analyse du réseau social des utilisateurs parmi les deux proposées ;
- Mener une évaluation comparative de notre approche relativement à une approche de recherche d'information classique.

4.2 Cadre d'évaluation

Le but est de mesurer l'apport des réseaux sociaux sur la qualité des résultats retournés par le moteur de recherche.

Les compagnes d'évaluations tel que TREC proposent un cadre standard pour évaluer et comparer les systèmes de recherche d'information. Cependant, les collections disponibles ne sont pas adaptées pour évaluer les modèles de recherche d'informations sociale car elles ne contiennent pas d'informations sociales.

L'évaluation d'un système de RI repose sur quatre éléments : un corpus de documents (base documentaire), un corpus de requêtes (liste de requêtes prédéfinies), des jugements de pertinence indiquant que tel document est pertinent pour telle requête, et des métriques d'évaluation.

4.2.1 Corpus de données (Collection de données)

Avec l'absence d'un cadre standard d'évaluation en recherche d'information sociale, et afin de valider notre proposition basée sur les réseaux sociaux, nous avons construit un corpus de données à partir de la collection Wiki10+[7]. Notre corpus contient les articles wiki annotés des mois de Janvier, Février, Mars et Avril 2009.

Wiki10 a été créée en Avril 2009 à partir du site de social bookmaking Del.icio.us et de Wikipédia.

En raison de l'absence d'informations sur l'identification des utilisateurs, cet ensemble est enrichi avec les données collectées depuis l'ensemble SocialBM0311[8]. Nous regroupons tous les identifiants des utilisateurs qui ont annoté les documents de la collection wiki10+ pendant les quatre premiers mois de l'année 2009[9].

Après filtrage, notre base contient **16608** articles et **229971** signets.

Les caractéristiques du réseau social des utilisateurs construit en utilisant la méthode 1 (i.e. sans prise en compte des tags) sont indiquées par le tableau 4.1.

Catégorie	Nombre
Nombre d'utilisateurs	**32529**
Nombre de relations	**13311208**

Tableau 4.1 : Caractéristiques du réseau social (sans prise en compte des tags)

Le tableau 4.2 contient les caractéristiques du réseau social des utilisateurs construit en utilisant la méthode 2 (i.e. avec prise en compte des tags).

[7] Wiki10+(Arkaitz Zubiaga. Enhancing Navigation on Wikipedia with Social Tags. Wikimania 2009. Buenos Aires, Argentina. 2009, http://nlp.uned.es/social-tagging/wiki10+/)

[8] (Arkaitz Zubiaga, Victor Fresno, Raquel Martinez, Alberto Perez Garcia-Plaza, *Harnessing Folksonomies to Produce a Social Classification of Resources*, IEEE Transactions on Knowledge and Data Engineering, vol. 25, no. 8, pp. 1801-1813, Aug. 2013, doi:10.1109/TKDE.2012.115**)**

[9] Le grand nombre de relations ainsi que la limite du logiciel VISONE qui a demandé une mémoire d'exécution dépassant 5 Go ont limité notre collection.

Catégorie	Nombre
Nombre d'utilisateurs	**25161**
Nombre de relations	**5458918**

Tableau 4.2 : Caractéristiques du réseau social (avec prise en compte des tags)

4.2.2 Corpus de requêtes (Collection de requêtes)

Les requêtes représentent un besoin des utilisateurs en information, et les tags sont des mots clés générés par les utilisateurs dans le but d'annoter les documents. Les tags peuvent ainsi être considérés comme des requêtes dans notre cadre d'évaluation. Nous considérons que les tags les plus populaires (i.e. les plus utilisés) sont de haute importance sociale et nous les sélectionnons comme requêtes.

Les **25** tags sélectionnés comme requêtes sont : art, books, culture, design, development, economics, history, imported, language, math, music, people, philosophy, politics, programming, psychology, reference, research, science, software, technology, web, web2, wiki, Wikipedia.

Les 25 requêtes sont soumises au moteur de recherche LUCENE et les 20 premiers résultats sont retenus et archivés pour chaque requête. Au total, 500 articles sont récupérés (25 requêtes × 20 résultats) et organisés sous forme de triplets <requête, url, score>. Le choix des 20 premiers résultats est justifié par le fait que ces derniers représentent les liens qui sont généralement visités par l'utilisateur sur l'ensemble des résultats retourné. Ils sont donc ceux qui contiennent les réponses les plus pertinentes. Néanmoins, nous signalons dans ce contexte, que ce nombre peut être élargi pour couvrir tous les résultats retournés.

4.2.3 Jugements de pertinence

Pour constituer la collection des documents pertinents, nous avons supposé qu'un document est pertinent s'il est annoté au moins une fois par le tag (requête).

La collection finale contient 25 requêtes et 286 documents pertinents avec une moyenne de 11,44 (=286/25) documents pertinents par requête.

4.2.4 Métriques d'évaluation

Dans ce cadre d'évaluation, nous utilisons la mesure classique de la précision pour les X premiers documents restitués (précision@X). C'est la proportion de documents pertinents dans les X premiers documents retrouvés pour chaque requête. Elle permet d'exprimer la satisfaction de l'utilisateur vis-à-vis des X premiers résultats pertinents. Elle constitue ainsi une mesure importante pour l'évaluation de la haute précision. Nous retenons les précisions pour les 10 et 15 premiers documents notés respectivement P@10 et P@15. A la fin, des moyennes de toutes ces précisions (pour chaque X=10 et 15) sont calculées sur toutes les requêtes de test.

Afin d'étudier l'impact du score social et de la combinaison du score social avec le score classique sur l'estimation globale de la pertinence, nous commençons par comparer différentes mesures de centralité pour choisir celle qui donne les meilleurs résultats.

4.3 Comparaison des mesures de centralité

Les utilisateurs sont principalement intéressés par les meilleurs résultats. La plupart d'entre eux examinent seulement les 20 premiers documents avant de prendre une décision [102].

Nous nous intéressons au jugement de pertinence pour le 1er résultat retourné par le moteur de recherche (P@1). Ce dernier a une importance particulière, puisque c'est le lien le plus cliqué par les utilisateurs. Nous étudions également les précisions pour les 10 et 15 premiers documents retrouvés, notés respectivement P@10 et P@15. A chaque niveau de pertinence, une valeur de 0 ou 1 est attribuée. 0 correspondant à un document non pertinent et 1 correspond à un document pertinent.

Nous appliquons les mesures de centralité suivantes : *Betweenness*, *Closeness*, et *PageRank* en utilisant les trois différentes techniques de pondération des liens entre utilisateurs.

Nous noterons l'application de ces mesures sur le modèle non pondéré par Betweenness0, Closeness0 et PageRank0 respectivement ; sur le réseau non-orienté pondéré par Betwenness1, Closeness1 et PageRank1 respectivement et sur le réseau orienté pondéré par Betwenness2, Closeness2 et PageRank2 respectivement.

Pour comparer les mesures d'importance sociale, les tableaux 4.3, 4.4 et 4.5 présentent pour la méthode 1 (sans prise en compte des tags) les valeurs obtenues par le classement des articles en utilisant uniquement les scores sociaux de leurs utilisateurs respectifs.

Binaire	Betweenness0	Closeness0	PageRank0
p@10	**0,4405**	0,0331	0,0842
p@15	**0,2984**	0,0238	0,0590

Tableau 4.3 : Classement en utilisant uniquement les scores sociaux (réseau non pondéré sans prise en compte des tags)

Pondéré	Betweenness1	Closeness1	PageRank1
p@10	0,0029	0,0356	**0,0356**
p@15	0,0020	0,0253	**0,0259**

Tableau 4.4 : Classement en utilisant uniquement les scores sociaux (réseau non-orienté pondéré sans prise en compte des tags)

Pondéré	Betweenness2	Closeness2	PageRank2
p@10	0,0752	0,0595	**0,0854**
p@15	0,0521	0,0416	**0,0597**

Tableau 4.5 : Classement en utilisant uniquement les scores sociaux (réseau orienté pondéré sans prise en compte des tags)

Les tableaux 4.6, 4.7 et 4.8 présentent pour la méthode 2 (avec prise en compte des tags) les valeurs obtenues par le classement des articles en utilisant uniquement les scores sociaux de leurs utilisateurs respectifs.

Binaire	Betweenness0	Closeness0	PageRank0
p@10	**0,3537**	0,0352	0,0782
p@15	**0,2372**	0,0246	0,0536

Tableau 4.6 : Classement en utilisant uniquement les scores sociaux (réseau non pondéré avec prise en compte des tags)

Pondéré	Betweenness1	Closeness1	PageRank1
p@10	0,0082	0,0371	**0,0381**
p@15	0,0055	0,0257	**0,0268**

Tableau 4.7 : Classement en utilisant uniquement les scores sociaux (réseau non-orienté pondéré avec prise en compte des tags)

Pondéré	Betweenness2	Closeness2	PageRank2
p@10	0,0567	0,0535	**0,0803**
p@15	0,0384	0,0367	**0,0550**

Tableau 4.8 : Classement en utilisant uniquement les scores sociaux (réseau orienté pondéré avec prise en compte des tags)

En comparant les précisions p@10 et p@15, nous constatons que la mesure de *Betweenness* appliquée sur le réseau non pondéré et la mesure de *PageRank* appliquée sur le réseau pondéré permettent de mieux classer les documents retournées initialement par un modèle de recherche d'information classique.

La pondération des liens du réseau social des utilisateurs permet d'améliorer l'efficacité de la recherche pour la mesure *Closeness* et *PageRank*. Cela est constaté avec les valeurs des précisions obtenues (voir tableau 4.5 et tableau 4.8) dépassant leurs analogues du réseau non pondéré (voir tableau 4.3 et tableau 4.6).

Pour évaluer l'efficacité de notre modèle dans ce qui suit, nous retenons la mesure de *Betweenness* pour le réseau non pondéré et la mesure de *PageRank* pour le réseau orienté pondéré comme étant les mesures qui permettent d'exprimer le mieux l'importance sociale des documents.

4.4 Evaluation de l'efficacité de l'approche proposée

Afin d'étudier l'importance du réseau social de l'utilisateur et sa capacité à exprimer la pertinence sociale des documents, nous comparons les résultats de notre modèle aux résultats du modèle vectoriel basé sur la mesure de classement TF-IDF**.**

4.4.1 Evaluation de l'efficacité de l'approche pour le réseau social sans prise en compte des tags

Nous comparons notre modèle de RIS proposé ainsi que le modèle social au modèle de recherche d'information classique.

a) Réseau non pondéré

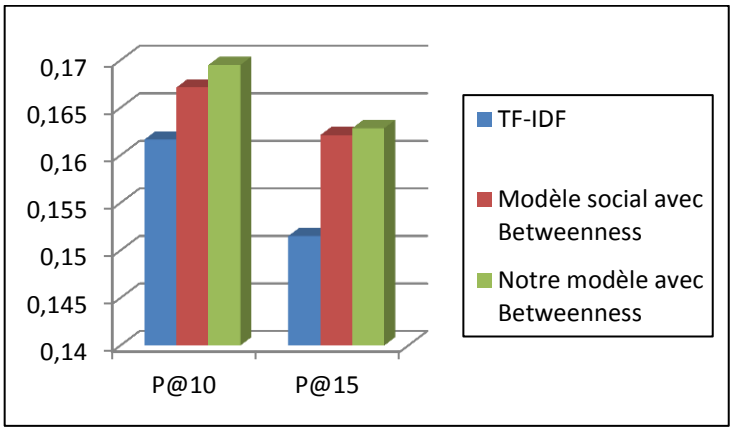

Figure 4.1 : Évaluation de l'efficacité de l'approche (utilisant le réseau non pondéré sans prise en compte des tags)

	TF-IDF	Modèle social	Notre modèle	Apport du modèle social % à TF-IDF	Apport de notre modèle % à TF-IDF
P@10	0,1616588	0,167134	0,1695036	3,39%	4,85%
P@15	0,1515123	0,162124	0,1628584	**7,00%**	**7,49%**

Tableau 4.9 : Apport du modèle social et du modèle combiné (utilisant le réseau non pondéré sans prise en compte des tags)

Comme décrit dans le tableau 4.9, notre modèle permet d'aboutir à une amélioration de la P@15 jusqu'à 7,49% par rapport au modèle classique.

De même, les résultats de l'importance sociale des documents améliorent la P@15 de 7% par rapport à TF-IDF.

b) Réseau pondéré

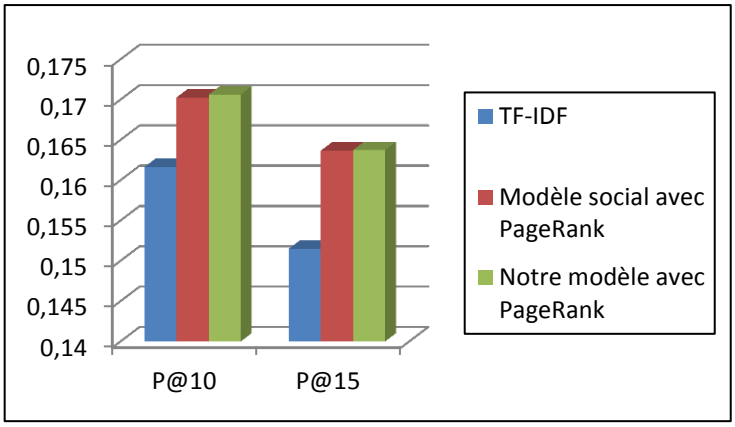

Figure 4.2 : Évaluation de l'efficacité de l'approche (utilisant le réseau pondéré sans prise en compte des tags)

	TF-IDF	Modèle social	Notre modèle	Apport du modèle social % à TF-IDF	Apport de notre modèle % à TF-IDF
P@10	0,1616588	0,1702828	0,1706248	5,33%	5,55%
P@15	0,15151227	0,1636816	0,16374053	**8,03%**	**8,07%**

Tableau 4.10 : Apport du modèle social et du modèle combiné (utilisant le réseau pondéré sans prise en compte des tags)

La combinaison des deux scores permet d'améliorer l'ordonnancement final des documents, la P@15 permet d'aboutir à une amélioration jusqu'à 8,07% par rapport au modèle classique.

Le score social permet d'améliorer l'efficacité de la recherche et la P@15 permet d'aboutir à un taux d'amélioration de 8,03% par rapport à TF-IDF.

Dans notre modèle, nous remarquons que les valeurs des précisions obtenues pour le réseau pondéré (voir tableau 4.10) dépassent leurs analogues pour le réseau non pondéré (voir tableau 4.9).

Dans modèle social, la P@15 permet une amélioration 8,03% pour le réseau pondéré comparé à 7% pour le réseau non pondéré.

4.4.2 Evaluation de l'efficacité de l'approche pour réseau social avec prise en compte des tags

Nous comparons notre modèle de recherche proposé ainsi que le modèle social au modèle vectoriel TF-IDF.

a) Réseau non pondéré

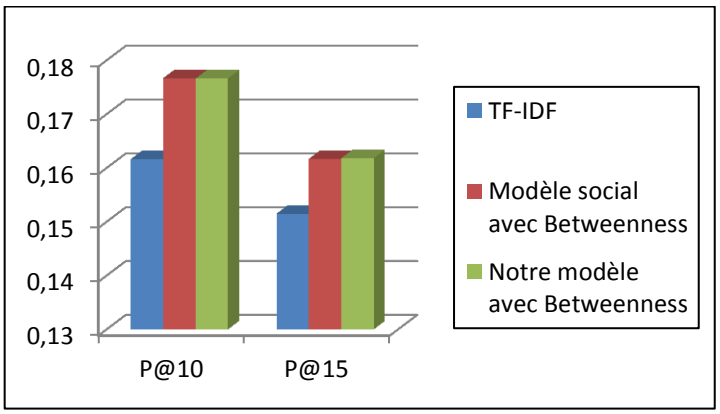

Figure 4.3 : Évaluation de l'efficacité de l'approche (utilisant le réseau non pondéré avec prise en compte des tags)

	TF-IDF	Modèle social	Notre modèle	Apport du modèle social % à TF-IDF	Apport de notre modèle % à TF-IDF
P@10	0,1616588	0,17671	0,1766468	**9,31%**	**9,27%**
P@15	0,15151227	0,1617184	0,16182053	6,74%	6,80%

Tableau 4.11 : Apport du modèle social et du modèle combiné (utilisant le réseau non pondéré avec prise en compte des tags)

Comme décrit dans le tableau 4.11, les meilleures valeurs de notre modèle permettent d'aboutir à une amélioration jusqu'à 9,27% par rapport au modèle classique.

Le modèle social permet d'améliorer la P@10 jusqu'à 9,31% par rapport à TF-IDF.

b) Réseau pondéré

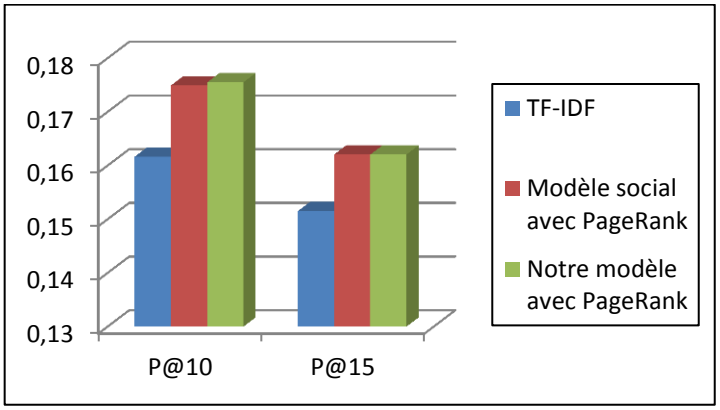

Figure 4.4 : Évaluation de l'efficacité de l'approche (utilisant le réseau pondéré avec prise en compte des tags)

	TF-IDF	Modèle social	Notre modèle	Apport du modèle social % à TF-IDF	Apport de notre modèle % à TF-IDF
P@10	0,1616588	0,1749324	0,1755048	**8,21%**	**8,56%**
P@15	0,15151227	0,1620504	0,1620504	6,96%	6,96%

Tableau 4.12 : Apport du modèle social et du modèle combiné (utilisant le réseau pondéré avec prise en compte des tags)

Les valeurs des précisions obtenues par notre modèle (voir tableau 4.12) dépassant celles du modèle classique.

Les résultats de l'importance sociale des documents améliorent la P@10 de 8,21% par rapport à TF-IDF.

Nous concluons donc que le choix de la formule de combinaison de la pertinence classique et l'importance sociale des documents permet d'améliorer l'efficacité de la recherche.

Nous remarquons que la P@15 donne un résultat meilleur pour le réseau pondéré (voir tableau 4.12) par rapport au réseau non pondéré (voir tableau 4.11).

4.4.3 Comparaison entre les deux méthodes

a) Réseau non pondéré

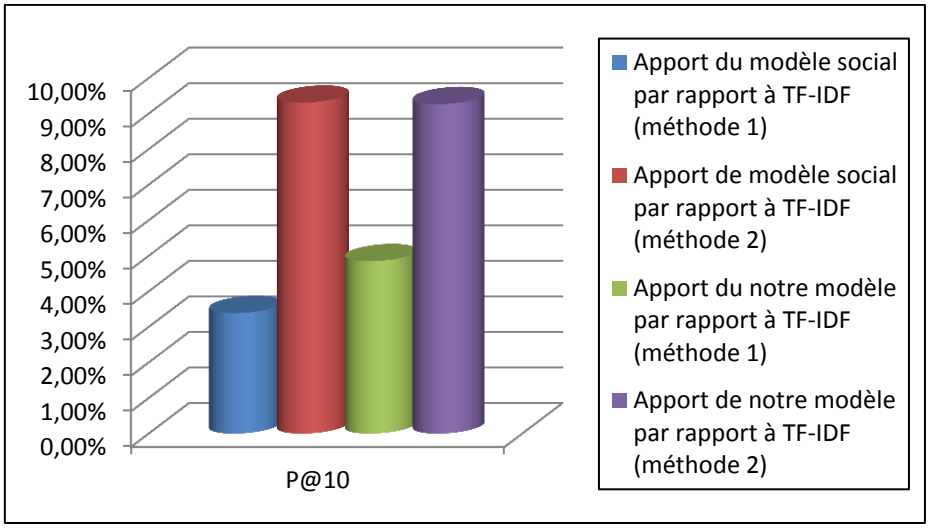

Figure 4.5 : Comparaison entre les deux méthodes (l'apport du modèle social et du modèle combiné par rapport au modèle classique en utilisant le réseau non pondéré)

	Apport du modèle social % à TF-IDF (utilisant la méthode 1)	Apport du modèle social % à TF-IDF (utilisant la méthode 2)	Apport du modèle social en utilisant la méthode 2 % à celui utilisant la méthode 1
P@10	3,39%	9,31%	**5,92%**

Tableau 4.11 : Apport du modèle social en utilisant la méthode 2 par rapport à celui utilisant la méthode 1 (réseau non pondéré)

	Apport de notre modèle % à TF-IDF (utilisant la méthode 1)	Apport de notre modèle % à TF-IDF (utilisant la méthode 2)	Apport du notre modèle en utilisant la méthode 2 % à celui utilisant la méthode 1
P@10	4,85%	9,27%	**4,42%**

Tableau 4.12 : Apport de notre modèle en utilisant la méthode 2 par rapport à celui utilisant la méthode 1 (réseau non pondéré)

Si nous considérons les dix (10) premiers résultats P@10, les valeurs sont remarquablement meilleures dans la méthode avec la prise en compte des tags. Notre modèle avec la méthode 2 obtient une amélioration de 9,27% et le modèle

avec la méthode 1 obtient une amélioration de 4,85% (comme décrit dans le tableau 4.12) ce qui donne un taux de 4,42% de plus.

Dans le modèle social (comme décrit dans le tableau 4.11), la précision pour le réseau avec prise en compte des tags permet d'améliorer l'efficacité de la recherche de 9,31% ; et pour le réseau sans prise en compte des tags obtient une amélioration de 3,39% ce qui donne un taux de 5,92% de plus.

b) Réseau pondéré

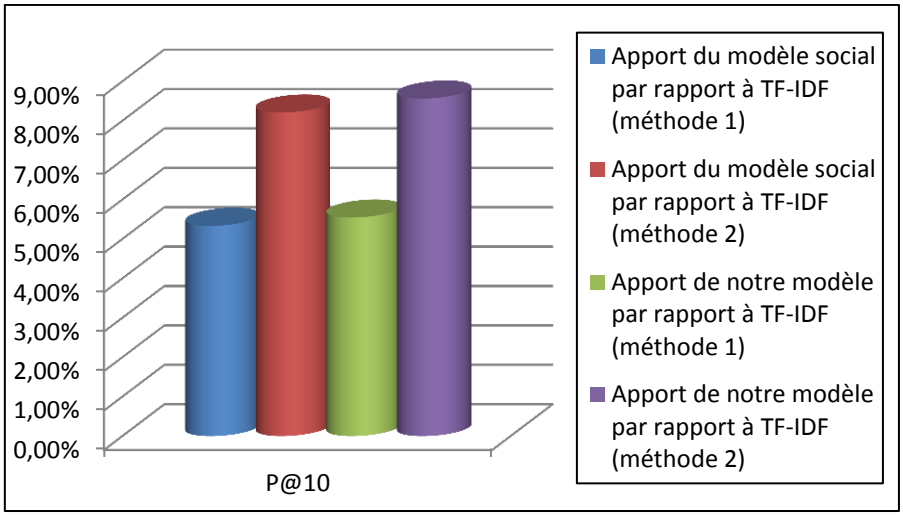

Figure 4.6 : Comparaison entre les deux méthodes (l'apport du modèle social et du modèle combiné par rapport au modèle classique en utilisant le réseau pondéré)

	Apport du modèle social % à TF-IDF (utilisant la méthode 1)	**Apport du modèle social % à TF-IDF (utilisant la méthode 2)**	**Apport du modèle social en utilisant la méthode 2 % à celui utilisant la méthode 1**
P@10	5,33%	8,21%	**2,88%**

Tableau 4.13 : Apport du modèle social en utilisant la méthode 2 par rapport à celui utilisant la méthode 1 (réseau pondéré)

	Apport de notre modèle % à TF-IDF (utilisant la méthode 1)	Apport de notre modèle % à TF-IDF (utilisant la méthode 2)	Apport du notre modèle en utilisant la méthode 2 % à celui utilisant la méthode 1
P@10	5,55%	8,56%	**3,02%**

Tableau 4.14 : Apport de notre modèle en utilisant la méthode 2 par rapport à celui utilisant la méthode 1 (réseau pondéré)

Si nous considérons les dix (10) premiers résultats P@10, Notre modèle avec la méthode 2 obtient une amélioration de 8,56% et le modèle avec la méthode 1 obtient une amélioration de 5,55% (comme décrit dans le tableau 4.14) ce qui donne un taux de 3,02% de plus.

Dans le modèle social, la précision pour le réseau avec prise en compte des tags permet d'améliorer l'efficacité de la recherche de 8,21%; et pour le réseau sans prise en compte des tags obtient une amélioration de 5,33% ce qui donne un taux de 2,88% de plus (comme décrit dans le tableau 4.13).

4.4.4 Evaluation de l'efficacité de l'approche pour le premier résultat

La situation est remarquablement meilleure si l'on ne considère que le premier résultat P@1, comme décrit dans la figure 4.7, notre modèle avec le premier réseau obtient une amélioration jusqu'à 33,38% et avec le second réseau une amélioration jusqu'à 34,28% par rapport à TF-IDF.

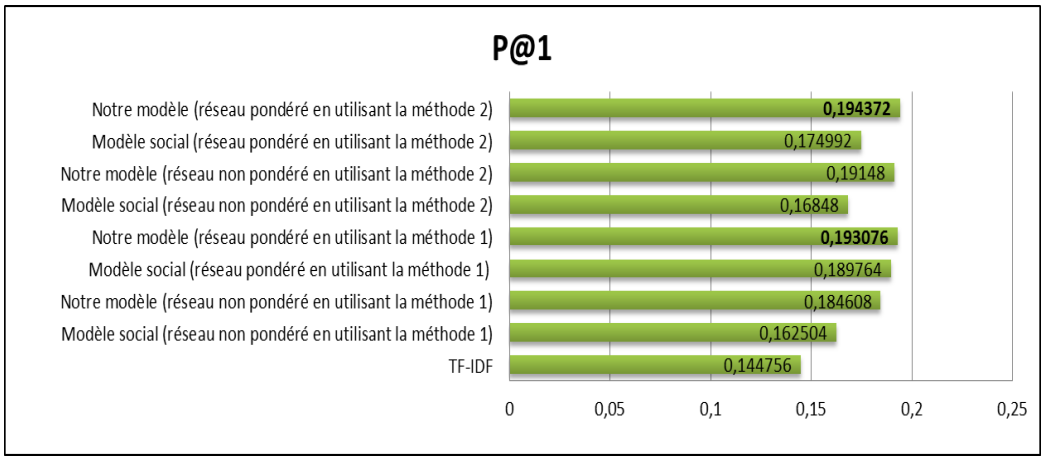

Figure 4.7 : Comparaison entre TF-IDF, les scores sociaux et les scores combinés avec les deux méthodes des deux réseaux (premier résultat)

4.5 Conclusion

Ce chapitre a présenté un cadre d'évaluation de notre approche qui intègre les relations sociales des utilisateurs du réseau de bookmarking ainsi que les tags utilisés pour marquer les documents et la pondération des relations sociales.

Les résultats expérimentaux ont montré en général l'efficacité de notre approche de recherche d'information sociale selon les deux modèles de réseaux sociaux, et en particulier :

1. la combinaison des deux dimensions de pertinences a amélioré l'ordonnancement final des documents comparativement à un système de recherche standard basé uniquement sur la pertinence classique ;
2. dans l'ensemble, la pondération des liens du réseau social a donné des résultats meilleurs par rapport au réseau binaire ;
3. pour le classement des dix (10) premiers résultat, notre approche a montré que la prise en compte des tags rend des résultats meilleurs comparés à ceux sans prise en compte des tags ;
4. la mesure de centralité de *PageRank* a permet de mieux évaluer l'importance sociale des documents ;
5. l'utilisation des réseaux sociaux a remarquablement amélioré la situation du premier résultat.

CONCLUSION

Les travaux présentés dans ce mémoire se situent dans le contexte de l'utilisation des réseaux sociaux dans les systèmes de recherche d'information et plus particulièrement dans le cadre du reclassement des résultats de recherche guidée par les informations sociales.

Nous avons proposé dans ce travail une approche permettant d'intégrer l'analyse des réseaux dans le processus de recherche d'information.

Dans notre approche, nous avons proposé un modèle de réseau d'information sociale qui intègre les documents, les annotations sociales et les utilisateurs. Nous avons présenté deux méthodes pour la construction du réseau social des utilisateurs : La première s'intéresse aux documents et aux utilisateurs et la seconde prend en plus en compte les tags.

Dans notre approche, nous avons proposé trois façons de pondération des liens entre les acteurs du réseau social. Une mesure de centralité est ensuite appliquée sur notre modèle de RIS pour le calcul de l'importance sociale des documents.

Nous avons combiné par la suite le score de l'importance sociale qui intègre la caractéristique sociale avec une valeur correspondant à la similarité entre la requête et le document pour améliorer le processus de recherche d'information.

Pour valider notre approche, nous l'avons évaluée en utilisant une collection extraite de la collection Wiki10+. Les résultats expérimentaux ont montré que la combinaison de la pertinence classique et l'importance sociale des documents permet d'améliorer l'efficacité de recherche.

En perspective, nous envisageons d'intégrer d'autres propriétés sociales au modèle proposé et de faire l'évaluation sur une grande collection de test.

Une autre poursuite de ce travail qui pourrait être intéressante serait d'intégrer la dimension sémantique à l'aide d'un thésaurus ou une ontologie comme WordNet pour la détection des relations entre les termes (les tags).

ANNEXE A
INTRODUCTION AU SYSTEME LUCENE

Lucene est un moteur de recherche et d'indexation développé dans le projet Apache. C'est un logiciel open source. A la base, Lucene est écrit en Java mais il est maintenant disponible pour d'autres langages de programmation tels que Python, PHP, Delphi, Perl, C++, … Lucene peut être utilisé avec de nombreux systèmes, c'est une multiplateforme du langage : Windows, Mac OS et Linux. Il est capable de traiter de grands volumes de documents grâce à sa puissance et à sa rapidité dues à l'indexation.

Calcul des scores en recherche de l'information

Le calcul de scores est propre au modèle vectoriel qui attribue à chaque requête un score de pertinence pour un document. Tf-Idf signifie Term frequency - Inverse document frequency.

- **Petit rappel des différents calculs**

$$\mathrm{idf}_t = \log \frac{N}{\mathrm{df}_t}$$

où N est le nombre de documents de la collection et df_t = nombre de documents dans lequel le terme apparaît.

$\mathrm{tf}_{t,d}$ = nombre d'occurrences du terme dans le document

$$\text{tf-idf}_t = \mathrm{tf}_{t,d} \times \mathrm{idf}_t$$

- **Calcul des scores avec Lucene**

La formule de calcul de score de Lucene[10] s'avère être compliquée :

score (q,d) = coord (q,d) × queryNorm (q) × Σ (tf(t in d) × idf(t) × t.getBoost() × norm(t,d))

[10] Lucene in Action chapitre 3.3 sur la compréhension des scores avec Lucene

où : **tf** est la racine carrée du tf usuel soit $\boldsymbol{tf}$**(t in d)** =$\sqrt{\text{tf}}$

$\boldsymbol{idf}$(t) =1 + log ($N \div df$ + 1)

coord (q,d) est le score calculé en fonction du nombre d'apparitions du document d dans la requête q. Ce type de score est issu du modèle booléen et est spécifique à Lucene.

queryNorm (q) est égal à la somme des carrés du poids de la requête et se calcule de cette façon : 1 ÷ $\sqrt{w^2}$ + w^2 + w^2

t.getBoost() = Boost est attribué au cours de l'indexation, par défaut le Boost est de 1. Il pourra être modifié par l'utilisateur au cours des requêtes évoquées

norm (t,d) qui englobe un coefficient d'importance des mots, documents et longueur des documents.

Exemples :

Pour une requête simple par terme, le score n'est calculé qu'à partir du « QueryWeight » qui est le poids de la requête effectuée.

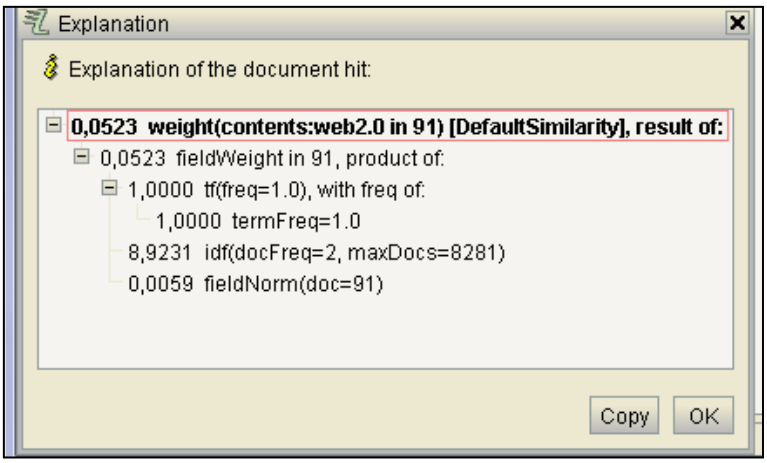

Figure A.1 : Explication d'un résultat à une requête simple

Si l'on prend le terme « web2.0 », l'outil résultat retourne 2 documents où l'on retrouve ce terme. Prenons l'exemple d'un de ses documents portant l'identifiant 91. Le score retourné est de 0.0523. Ce score est calculé à partir d'un seul terme, et d'un seul champ qui porte un poids. Le poids du champ est égal au score retourné. Pour retrouver le score, on doit faire ce calcul en utilisant la formule :

fieldWeight = tf (t) × idf (t,d) × fieldNorm.

Pour une requête à plusieurs termes, le score obtenu est calculé en faisant la somme des poids de chaque terme.

Le poids est calculé de cette façon :

Weight =Σ (QueryWeight × FieldWeight).

QueryWeight = idf × QueryNorm

FieldWeight = tf × idf × FieldNorm

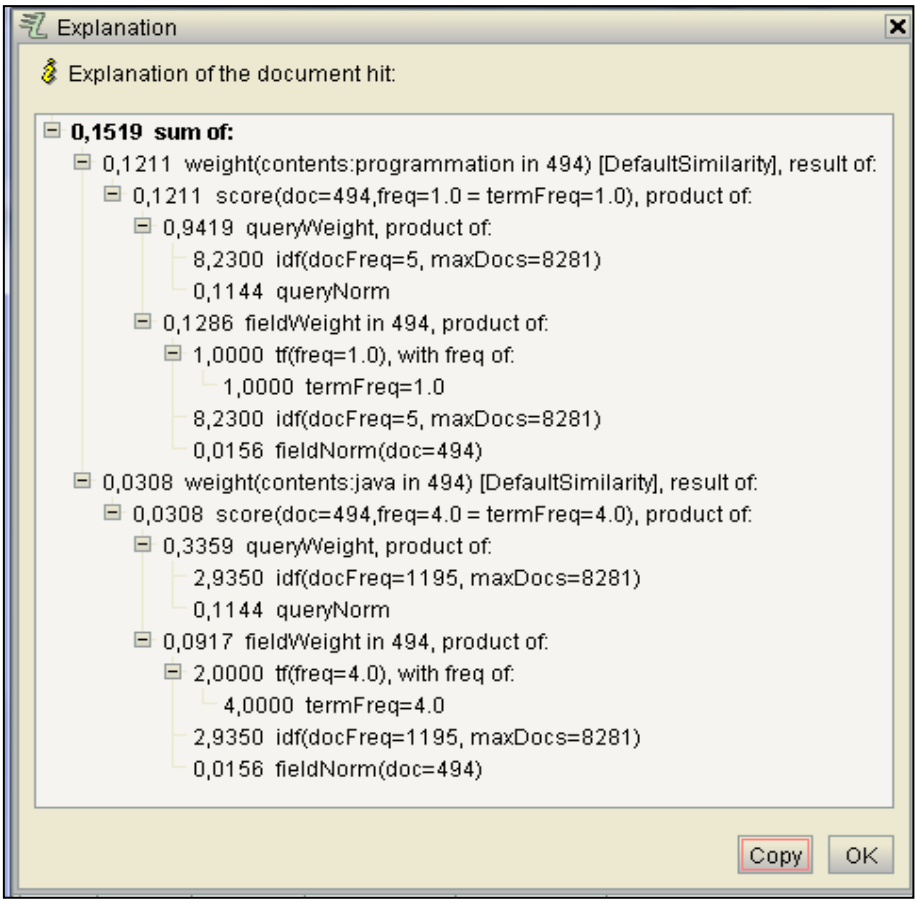

Figure A.2 : Explication d'un résultat à une requête composée

Si l'on prend la requête « programmation AND java », on obtient ce type de score (voir figure A.2). Lucene fait la somme du poids du terme « programmation » et de celui du terme « java ». Ces deux termes font tous les deux partis du champ : **contents**. « 494 » correspond à l'identifiant unique d'un des documents.

ANNEXE B
COLLECTION DE DONNEES

Wiki10+ est un ensemble de données constitué de 20764 articles en anglais de *Wikipédia*, avec leurs étiquettes sociales correspondantes. Chacun d'entre eux est annoté par au moins 10 utilisateurs sur *Del.icio.us.*

Deux types de données sont présents dans cet ensemble :

- les étiquettes[11] utilisées pour le marquage des articles de *Wikipédia* ;
- les articles[12] de *Wikipedia*.

Le format des métadonnées

L'ensemble de données sur les documents (les métadonnées) est fourni au format XML, suivant ce modèle :

```
<articles>
...
<article>
<hash>MD5 hash for document's URL</hash>
<title>The title of the article</title>
<users>Number of users annotating it</users>
<tags>
...
<tag>
<name>Tag name</name>
<count># of users who annotated the tag</count>
</tag>
...
</tags>
</article>
```

[11] *wiki10 + _tag-data.tar.gz*

[12] *documents.tar.bz2 wiki10 +*

...
</article>

SocialBM0311 est une grande collection d'étiquetage de données recueillies auprès de *Del.icio.us*[13]. Elle contient toute l'activité de *bookmarking* pour près de 2 millions d'utilisateurs du lancement du site *Del.icio.us* en 2003 à la fin Mars 2011. La base de données contient :

- 339897227 signets ;
- 118520382 URL uniques ;
- 14723731 étiquettes uniques ;
- 1951207 utilisateurs.

Les fichiers contiennent un signet par ligne, avec les champs suivants séparés par des tabulations:

Url_md5 User_id Url Unix_timestamp Tags

où:

- « Url_md5 » est le hachage MD5 de l'URL du signet. *del.icio.us* utilise le hachage MD5 comme identifiant de l'URL.
- « User_id » est l'identifiant de l'utilisateur qui a enregistré le signet. Les noms d'utilisateurs sont entièrement anonymes pour cet ensemble de données, et les identifiants d'utilisateur fournies avec le jeu de données ont été assignés au hasard à des utilisateurs.
- « Url » est l'URL du signet.
- « Unix_timestamp » se réfère à la date à laquelle le signet a été enregistré, en utilisant le format standard de temps UNIX. Notez que ces horodateurs sont arrondis à jour, et ne fournissent pas le moment précis (limité par le système lors de la collecte de données).
- « Tags » comprennent une liste séparée par des tabulations des tags (mots-clés) utilisé dans le signet.

---

[13] *Delicious.com*

REFERENCES

1. **Salton, G., McGill M.** *"Introduction to Modern Information Retrieval". McGraw-Hill, New York.* 1983a.

2. **Hernandez, N.** *"Ontologie de domaine pour la modélisation du contexte en recherche d'information". Thèse de doctorat. Université de Toulouse Paul Sabatier.* 2006.

3. **Daoud, M.** *"Accès personnalisé à l'information : approche basée sur l'utilisation d'un profil utilisateur sémantique dérivé d'une ontologie de domaines à travers l'historique des sessions de recherche". Thèse de doctorat. Université Paul Sabatier de Toulouse.* 2009.

4. **Baziz, M.** *"Indexation conceptuelle guidée par ontologie pour la recherche d'information". Thèse de doctorat. Université Paul Sabatier de Toulouse.* 2005.

5. **Ingwersen, P.** *"Information retrieval interaction". London, Taylor Graham.* 1992.

6. **Mallak, I.** *"De nouveaux facteurs pour l'exploitation de la sémantique d'un texte en Recherche d'Information". Thèse de doctorat. Université Toulouse 3 Paul Sabatier.* 2011.

7. **Barry, C. L.** *"User-defined relevance criteria : an exploratory study". Journal of the American Society for Information Science.* 1994.

8. **Smeaton, A. F.** *"Information retrieval and natural language processing". In Proceedings of a conference jointly sponsored by ASLIB, university of York.* 1989.

9. **Tambellini, C.** *"Un système de recherche d'information adapté aux données incertaines : adaptation du modèle de langue". Thèse de doctorat. Université de Nice-Sophia Antipolis-UFR sciences.* 2007.

10. **Belkin, N., Croft, W.** *"Information Retrieval and Information Filtering: Two Sides of the same Coin". Communications of the ACM.* 1992.

11. **Amrouche, K.** *"Passage à l'échelle en Recherche d'Information : Méthode d'élagage pour la réduction de l'espace de recherche". Thèse de doctorat. INI.* 2008.

12. **Salton, G.** *"A comparison between manual and automatic indexing methods". Journal of American Documentation.* 1971.

13. **Robertson, S. E.** *"The probability ranking principle in IR". Journal of Documentation.* 1977.

14. **Kompaoré, N. D. Y.** *“Fusion de systèmes et analyse des caractéristiques linguistiques des requêtes : vers un processus de RI adaptatif». Thèse de doctorat. Université Paul Sabatier de Toulouse.* 2008.

15. **Boughanem, M., Chrisment, C.** *"Query modification based on relevance back-propagation in an ad hoc environment". Inf. Process. Manage.* 1999.

16. **Andreasen, T.** *"An approach to knowledge-based query evaluation : Fuzzy databases. Fuzzy Sets and Systems". Publisher : Elsevier Science.* 2003.

17. **Khan, L. R.** *"Ontology-based Information Selection". Phd thesis. University of Southern California.* 2000.

18. **Crouch, C. J., Yang, B.** *"Experiments in automatic statistical thesaurus construction". In Proceedings of 15th annual international ACM SIGIR Conference on Research and Development in Information Retrieval. ACM.* 1992.

19. **Hiemstra, D., Robertson, S.** *"Relevance feedback for best match term weighting algorithms in information retrieval". In A. Smeaton and J. Callan, editors, Proceedings of the Joint DELOS-NSF Workshop on Personalisation and Recommender Systems in Digital Libraries, ERCIM Workshop.* 2001.

20. **Baeza-Yates, R., Ribeiro-Neto, R. A.** *“Modern Information Retrieval”. New York : ACM Press ; Harlow England : Addison-Wesley, cop.* 1999.

21. **Bouramoul, A.** *"Recherche d'information contextuelle et sémantique sur le Web". Thèse de doctorat. Université Mantouri Constantine.* 2011.

22. **Salton, G., Fox, E.A., Wu, H.** *"Extended Boolean information retrieval system". Commun ACM.* 1983b.

23. **Zadeh, L. A.** *" Fuzzy sets. Information and Control". 8, 338 - 353.* 1965.

24. **Salton, G., McGill, M.** *"Introduction to Modern Information Retrieval". McGraw-Hill, New York.* 1983a.

25. **Deerwester, S., Dumais, S., Furnas, G., Landauer, T., Harshman, R.** *"Indexing by latex semantic analysis". Journal of the American Society for Information Science.* 1990.

26. **Wong, S., Ziarko, W., Wong, P.** *"Generalized vector space model information retrieval", In Proceedings of the 8th annual international ACM SIGIR Conference on Research and development in Information Retrieval, ACM.* 1985.

27. **Robertson, S. E., Sparck, J. K.** *"Relevance weighting for search terms". Journal of The American Society for Information Science.* 1976.

28. **Maron, M. E., Kuhns, J. L.** *"On relevance, probabilistic indexing and information retrieval". Journal ACM.* 1960.

29. **Tamine, L., Boughanem, M., Daoud, M.** *"Evaluation of contextual information retrieval effectiveness: overview of issues and research". Knowledge and Information Systems.* 2010.

30. **Ingwersen, P., Järvelin, K.** *"The TURN : Integration of Information Seeking and Retrieval in Context". SPRINGER.* 2005.

31. **Saracevic, T.** *"The stratefied model of information retrieval interaction: extension and applications". In Proceedings of the 60th annual meeting of the American Society for Information Science.* 1997.

32. **Calabretto, S., Egyd-Zsigmond, E.** *"Recherche d'Information en contexte". EARIA'06, France.* 2006.

33. **Allan, J., al.** *"Challenges in information retrieval and language modeling : report of a workshop held at the center for intelligent information retrieval". University of massachusetts amherst, september 2002. SIGIR Forum,.* 2003.

34. **Lang, K.** *"Newsweeder : learning to filter netnews". In Proceedings of the 12th International Conference on Machine Learning. Morgan Kaufmann publishers Inc. San Mateo, CA, USA.* 1995.

35. **Smyth, B., Balfe, E.** *"Anonymous personalization in collaborative web search. Information retrieval". Journal ACM.* 2006.

36. **Tamine, L.** *"De la recherche d'information orientée système vers la recherche d'information orientée contexte : Verrous, contributions et perspectives". Habilitation à Diriger des Recherches. Université Sablier, Toulouse.* 2008.

37. **Bouidghaghen, O.** *"Accès contextuel à l'information dans un environnement mobile : approche basée sur l'utilisation d'un profil situationnel de l'utilisateur et d'un profil de localisation des requêtes". Thèse de doctorat. Université Toulouse 3 Paul Sabatier.* 2011.

38. **Tanudjaja, F., Mui, L.** *"Persona : A contextualized and personalized web search". In proceedings 35th Hawaii International Conference on System Sciences.* 2002.

39. **Serir, F.** *"Annotation d'un corpus pour l'évaluation des systèmes de recherche d'informations", mémoire d'ingéniorat, université Abou bakr Belkaid - Tlemcen.* 2012.

40. **Lafrage, L.** *"Recensement des données disponibles via les campagnes d'évaluation des systèmes de recherche d'information". Rapport internet, Version 1, IRIT.* 2008.

41. **Hedin, S. Hoestlandt, M., Lenouvel, S., Verraest, S.** *"L'évaluation des systèmes de recherche d'informations". http://idemm.joueb.com/news/l-evaluation-des-systemes-de-recherche-d-informations.* 2004.

42. **Kleinberg, J.** *"The convergence of social and technological networks". Commun. ACM.* 2008.

43. **Ben jabeur, L.** *"Leveraging social relevance: Using social networks to enhance literature access and microblog search". Thèse de doctorat. Université Toulouse 3 Paul Sabatier.* 2013.

44. **Passant, A.** *"Technologies du Web Sémantique pour l'Entreprise 2.0". Thèse de doctorat. Université Paris IV - Sorbonne.* 2009.

45. **Stankovic, M.** *"Convergence entre Web Social et Web Sémantique : Application à l'innovation à l'aide du Web". Thèse de doctorat. Université Parie - Sorbonne.* 2012.

46. **Mlaiki, A.** *"Compréhension de la continuité d'utilisation des réseaux sociaux numérique : les apports de la théorie du don". Thèse de doctorat. Paris - DALPHINE.* 2012.

47. **Leuf, B., Cunningham, W.** *"The Wiki Way: Collaboration and Sharing on the Internet". Addison-Wesley Professional.* 2001.

48. **Malek, M.** *"Introduction à l'analyse des réseaux sociaux". http://mma.perso.eisti.fr/HTML-IAD/Seminaire1.pdf.* 2009.

49. **Bouhini, C., Géry, M., Largeron, C.** *"Modèle de Recherche d'Information Sociale Centré Utilisateur". EGC : 275-286.* 2013.

50. **Kleinberg, J. M., Kumar, R., Raghavan, P., Rajagopalan, S., Tomkins, A. S.** *"The Web as a Graph: Measurements, Models, and Methods". Book Series Lecture Notes in Computer Science. Editeur Springer Berlin / Heidelberg.* 1999b.

51. **Mutschke, P.** *"Enhancing Information Retrieval in Federated Bibliographic Data Sources Using Author Network Based Stratagems". SpringerLink . Book Research and Advanced Technology for Digital Libraries.* 2001.

52. **Krebs, V.** *"The Social Life of Routers: Applying Knowledge of Human Networks to Design of Computer Networks". The Internet Protocol Journal.* 2000.

53. **Merckle, P.** *"Les origines de l'analyse des réseaux sociaux". Article produit dans le cadre d'un cours pour l'Ecole Nationale Supérieure de Lyon en 2003-2004. .* 2003.

54. **Chikhi, N. F.** *"Calcul de centralité et identification de structures de communautés dans les graphes de documents". Thèse de doctorat. Université Toulouse 3 Paul Sabatier.* 2010.

55. **Freeman, L. C.** *"Centrality in Social Networks : Conceptual Clarification". Social Networks.* 1979.

56. **Brin, S., Page, L.** *"The anatomy of a large-scale hypertextual Web search engine". Computer Network and ISDN Systems.* 1998.

57. **Zhang, Y., Yu, J. X., Hou, J.** *"Web Communities: Analysis and Construction". Springer.* 2005.

58. **Kleinberg, J.** *"Authoritative sources in a hyperlinked environment". In Proceedings of the ninth annual ACM-SIAM symposium on Discrete algorithms, San Francisco, California, United States, Society for Industrial and Applied Mathematics.* 1998.

59. **Kleinberg, J**. *" Authoritative sources in a hyperlinked environment". Journal ACM.* 1999a.

60. **TA, T. A.** *"Web sémantique et réseaux sociaux - Construction d'une mémoire collective par recommandations mutuelles et (re-)présentations". Thèse de doctorat. Ecole Nationnale Supérieure des Télécommunications, Paris.* 2005.

61. **Kirsch, S., Gnasa, M., Cremers, A.** *"Beyond the web: Retrieval in social information spaces". Advances in Information Retrieval.* 2006.

62. **Amer-Yahia, S., Benedikt, M., Bohannon, P.** *"Challenges in searching online communities". IEEE Data Eng. Bull.* 2007.

63. **Ben Jabeur, L., Tamine, L., Boughanem, M.** *" A social model for literature access : towards a weighted social network of authors. In Proceedings of the 9th conference Recherche d'Information Assistée par Ordinateur, RIAO'10.* 2010.

64. **Kirchhoff, L., Stanoevska-Slabeva, K., Nicolai, T., Fleck, M.** *"Using social network analysis to enhance information retrieval systems". Technical report. University of St.Gallen - Alexandria Repository (Switzerland).* 2008.

65. **Siersdorfer, S., Sizov, S.** *"Social recommender systems for web 2.0 folksonomies". 20th ACM Conference on Hypertext and Hypermedia, Hypertext.* 2009.

66. **Tang, J., Zhang, J., Yao, L., Li, J., Zhang, L., Su, Z.** *"Arnetminer: extraction and mining of academic social networks". In Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining, KDD '08, New York, NY, USA. ACM.* 2008.

67. **Schifanella, R., Barrat, A., Cattuto, C., Markines, B., Menczer, F.** *"Folks in folksonomies: social link prediction from shared metadata". In Proceedings of the third ACM international conference on Web search and data mining, WSDM '10, New York, NY, USA. ACM.* 2010.

68. **Zhang, Y.** *"Learning, innovating, and an emerging core of knowledge: A model for the growth of citation networks". Available at SSRN 1975606.* 2011a.

69. **Newman, M. E. J., Girvan, M.** *"Finding and evaluating community structure in networks". Physical Review E 69, 026113.* 2004.

70. **Crandall, D., Backstrom, L., Cosley, D., Suri, S., Huttenlocher, D., Kleinberg, J.** *"Inferring social ties from geographic coincidences". In Proceedings of the National Academy of Sciences.* 2010.

71. **Song, X., Chi, Y., Hino, K., Tseng, B.** *"Identifying opinion leaders in the blogosphere". In Proceedings of the sixteenth ACM conference on Conference on information and knowledge management, CIKM '07, New York, NY, USA. ACM.* 2007.

72. **Bodendorf, F., Kaiser, C.** *" Detecting opinion leaders and trends in online social networks". In Proceedings of the 2nd ACM workshop on Social web search and mining. ACM.* 2009.

73. **Korfiatis, N. T., Poulos, M., Bokos, G.** *"Evaluating authoritative sources using social networks: an insight from wikipedia". Online Information Review.* 2006.

74. **Bouadjenek, M. R.** *"Infrastructure and Algorithms for Information Retrieval Based On Social Network analysis/Mining". PhD thesis. PRiSM Laboratory, Versailles University.* 2013.

75. **Morris, M. R., Teevan, J., Panovich, K.** *"What do People Ask Their Social Networks, and Why?: A Survey Study of Status Message Q&A Behavior". In Proceedings of the 28th international conference on Human factors in computing systems, CHI '10, New York, USA, ACM.* 2010.

76. **Ma, H., Yang, H., Lyu, M. R., Kingand, I.** *"SoRec: social recommendation using probabilistic matrix factorization". In Proceedings of the 17th ACM conference on Information and knowledge management, CIKM '08, , New York, NY, USA, ACM.* 2008.

77. **Wen, K., Li, R., Xia, J., Gu, X.** *"Optimizing ranking method using social annotations based on language model". Artificial Intelligence Review.* 2012.

78. **Vallet, D., Cantador, I., Jose, J. M.** *"Personalizing web search with folksonomy-based user and document profiles". In Proceedings of the 32nd European Conference on Advances in Information Retrieval, ECIR'10, Springer Berlin Heidelberg.* 2010.

79. **Kirsch, S. M.** *"Social Information Retrieval". Phd thesis. Rheinische Friedrich-Wilhelms-Universitat Bonn.* 2005.

80. **Bischoff, K., Claudiu, Firan, C. S., Nejdl, W., Paiu, R.** *" Can all tags be used for search?". In Proceedings of the 17th ACM conference on Information and knowledge management, CIKM '08, New York, NY, USA, ACM.* 2008.

81. **Dror, G., Koren, Y., Maarek, Y., Szpektor, I.** *"I want to answer; who has a question?: Yahoo! answers recommender systems". In Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining, KDD '11, New York, NY, USA. ACM.* 2011.

82. **Horowitz, D., Kamvar, S. D.** *"The anatomy of a large-scale social search engine". In Proceedings of the 19th international conference on World wide web, WWW '10, New York, NY, USA. ACM.* 2010.

83. **Filho, F. F., Gary, M. O., Geus, P. L.** *"Kolline: a task-oriented system for collaborative information seeking". In Proceedings of the 28th ACM International Conference on Design of Communication, SIGDOC '10, New York, NY, USA. ACM.* 2010.

84. **Jansen, B. J., Chowdury, A., Cook, G.** *"The ubiquitous and increasingly significant status message. interactions". Journal ACM.* 2010.

85. **Guy, I., Jacovi, M., Shahar, E., Meshulam, N., Soroka, V., Farrell, S.** *"Harvesting with SONAR: the value of aggregating social network information". In CHI. Computer Human Interaction.* 2008.

86. **Guy, I., Zwerdling, N., Ronen, I., Carmel, D., Uzizl, E.** *"Social media recommendation based on people and tags". Special Interests in Information Retrieval SIGIR.* 2010.

87. **Tchuente, D.** *" Modelisation et derivation de profils utilisateurs a partir de reseaux sociaux : approche a partir de communautes de reseaux k-egocentriques"». Thèse de doctorat. Universite de Toulouse 3 - Paul Sabatier- Toulouse.* 2013.

88. **Wang, X., Liu, H., Fan, W.** *"Connecting users with similar interests via tag network inference". In Proceedings of the 20th ACM international conference on Information and knowledge management, CIKM '11, New York, NY, USA. ACM.* 2011.

89. **Kautz, H., Selman, B., Shah, M.** *"Referral Web: combining social networks and collaborative filtering". Commun. ACM.* 1997.

90. **Agosto, L.** *"Optimisation d'un Réseau Social d'Échange d'Information par Recommandation de Mise en Relation". Thèse de doctorat. Université de Savoie.* 2005.

91. **Mika, P.** *"Ontologies Are Us: A Unified Model of Social Networks and Semantics". In Proceedings of the 4th International Semantic Web Conference (ISWC 2005), volume 3729 de Lecture Notes in Computer Science, Springer.* 2005.

92. **Zhang, Z. K., Zhou, T., Zhang, Y. C.** *"Tag-aware recommender systems: A state-of-the-art survey". Journal of Computer Science and Technology.* 2011b.

93. **Jelassi, M. N., Ben Yahia, S., Nguifo, E. N.** *"Vers des recommandations plus personnalisées dans les folksonomies". Actes Conference IC.* 2014.

94. **Ricci, F., Rokach, L., Shapira, B., Kantor, P.** *"Recommender Systems Handbook" . Springer.* 2011.

95. **Diederich, J., Iofciu, T.** *" Finding communities of practice from user profiles based on folksonomies". In Proceedings of the 1st International Workshop on TEL-CoPs, Crete, Greece.* 2006.

96. **Hu, J., Wang, B., Tao, Z.** *"Personalized tag recommendation using social contacts". In Proc. Of Workshop SRS'11, in conjunction with CSCW.* 2011.

97. **Jäschke, R., Marinho, L., Hotho, A., Schmidt-Thieme, L., Stumme, G.** *"Tag recommendations in folksonomies". In Procedding of the 11th ECML(PKDD), Warsaw, Poland.* 2007.

98. **Lipczak, M.** *"Tag recommendation for folksonomies oriented towards individual users". In proceeddings of the ECML/PKDD Discovery Challenge, Antwerp, Belgium.* 2008.

99. **Wasserman, S., Faust, K.** *"Social Network Analysis: Methods and Applications". Structural Analysis in the Social Sciences. Cambridge University Press.* 1994.

100. **Hatcher, E., Gospodnetic, O.** *"Lucene in Action". Manning Publications. ISBN 1.932394.28.1.* 2005.

101. **Baur, M.** *"VISONE, Software for the Analysis and Visualization of Social Networks". Fakultät für Informatik, Institut für Informatik Theoretische.* 2008.

102. **Spink, A., Wolfram, D., Jansen, M. B., Saracevic, T.** *"Searching the web: The public and their queries". Journal of the American society for information science and technology.* 2001.

103. **Xie, H. I.** *"Users' evaluation of digital libraries (dls): Their uses, their criteria, and their assessment". Information Processing and Management.* 2008.

104. **Rijsbergen, C.** *"Information Retrieval". Butterworth & Co (Publishers). Ltd, second edition, London.* 1979.

105. **Liu, F., Lee, H. J.** *"Use of social network information to enhance collaborative filtering performance". Journal ACM.* 2010.

106. **Kim, K.** *"Effects of emotion control and task on web searching behavior". Information Processing and Management.* 2008.

107. **Hupfer, M. E., Detlor, B.** *"Gender and web information seeking: A self-concept orientation model". Journal of the American Society for Information Science and Technology.* 2006.

108. **Fuhr, N.** *"Information retrieval : introduction and survey". Post-graduate course on information retrieval, university of Duisburg-Essen, Germany.* 2000.

109. **Cool, C.** *"The concept of situation in information science". Annual review of information science and technology.* 2001.

110. **Cheverst, K., Davies, N., Mitchell, K., Friday, A., Efstratiou, C.** *"Developing a context-aware electronic tourist guide : some issues and experiences". In Proceedings of the SIGCHI conference on Human factors in computing systems, New York, NY, USA, ACM.* 2002.

111. **Broder, A.** *"A taxonomy of web search". SIGIR Forum.* 2002.

112. **Bilal, D.** *"Children's use of the yahooligans ! web search engine : cognitive, physical, and affective behaviors on fact-based search tasks". Journal of the American Society for Information Science.* 2000.

113. **Ingwersen P.** *"Polyrepresentation of information needs and semantic entities: elements of a cognitive theory for information retrieval interaction". .* s.l. : In Proceedings of the Seventeenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval., 1994.

114. **Tombros, A., Ruthven, I., Jose, J. M.** *"How users assess web pages for information seeking". Journal of the American Society of Information Science and Technology (JASIST).* 2005.

115. **Timothy, M., Sherry, T., Robert, M.** *"Hypermedia learning and prior knowledge : domain expertise vs. system expertise". Journal of Computer Assisted Learning.* 2005.

116. **Smith, M., Barash, V., Getoor, L., Lauw, H. W.** *"Leveraging social context for searching social media". In SSM '08 : Proceeding of the 2008 ACM workshop on Search in social media, New York, NY, USA, ACM.* 2008.

117. **Kumar, R., Raghavan, P., Rajagopalan, S., Tomkins, A.** *"Core algorithms in the CLEVER system". ACM Trans. Internet Technol.* 2006.

118. **Frias-Martinez, E., Chen, S. Y., Macredie, R. D., Liu, X.** *"The role of human factors in stereotyping behavior and perception of digital library users : a robust clustering approach". User Modeling and User-Adapted Interaction.* 2007.