République Algérienne Démocratique et Populaire
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique
Université Mouloud Mammeri de Tizi-Ouzou

Faculté de : Génie électrique et d'informatique
Département : Informatique

Mémoire de fin d'études

En vue de l'obtention du diplôme de
Master en Informatique

Spécialité : Systèmes informatiques

Présenté par :
SEKOUR Mohamed

Thème

« Exploitation des signaux sociaux de Twitter pour améliorer la recherche d'information »

Proposé et dirigé par : Mme FELLAG Samia

Soutenu le 11 juillet 2019 devant le jury composé de :

Mme G. SINI	Présidente du Jury
Mme Y. YASLI	Membre du Jury
Mme S. FELLAG	Directrice de mémoire

Dédicaces

Je dédie ce modeste travail à mes très chers parents qui n'ont jamais cessé de m'encourager, qui ont toujours cru en moi et à tous leurs sacrifices.

À mes frères et sœurs pour leurs soutiens, à toute ma famille et à mes très chers amis qui sont restés auprès de moi tout au long de mon travail et qui m'ont donné le courage pour le réussir.

Remerciements

Je tiens à exprimer toute ma reconnaissance à ma directrice de mémoire, Madame Samia FELLAG. Je la remercie de m'avoir encadré, orienté, aidé et conseillé.

Mes remerciements s'adressent également aux membres du jury qui ont accepté d'évaluer mon travail.

Je tiens à exprimer mes sincères remerciements à tout le corps professoral et administratif de l'université Mouloud Mammeri de Tizi-Ouzou.

Enfin, je remercie tous ceux qui, de près ou de loin, ont contribué à la réalisation de ce travail.

Résumé

Notre travail se situe dans le contexte de la recherche d'information sociale (RIS) et s'intéresse plus particulièrement à l'exploitation des signaux sociaux dans le processus de la recherche d'information (RI). Les signaux sociaux représentent des informations communicatives et informatives qui fournissent directement ou indirectement des renseignements sur les interactions, les émotions, les relations et les comportements sociaux. Bien que la RI classique a pu faire face à la multiplication des données informatiques et leurs volumes considérables et cela en facilitant l'accès à ces informations en développant des systèmes de recherche d'information (SRI), la RI a évolué avec l'émergence du Web et plus récemment des réseaux sociaux (RS). De nos jours, les RS représentent le moyen le plus utilisé pour la communication, le partage de connaissance et de contenus sur le Web. Avec cette dimension sociale qui vient enrichir les contenus des ressources sur le Web, les utilisateurs se retrouvent avec de nouveaux besoins en information. D'où l'émergence de la RI Sociale (RIS), une thématique récente qui a pour objectif de prendre en compte les informations spécifiques aux RS. L'objectif principal de notre travail consiste à intégrer les différents signaux issus des réseaux sociaux dans le processus de la RI afin d'améliorer la recherche et de répondre aux besoins en informations de l'utilisateur. En effet, dans ce mémoire, nous proposons deux approches de RI sociale qui exploitent les signaux sociaux du réseau social *Twitter*, la première s'intéresse à l'indexations des *Tweets* et la deuxième à l'utilisation de ces *Tweets* pour reformuler la requête initiale.

Mots clés : Recherche d'information, recherche d'information sociale, réseaux sociaux, signaux sociaux, Twitter.

Abstract

Our work is in the context of social information retrieval (SIR) and is particularly interested in the exploitation of social signals in the process of information retrieval. Social signals represent communicative and informative information that directly or indirectly provides information about social interactions, emotions, relationships and behaviours. Although traditional IR has been able to cope with the proliferation of computer data and its considerable volumes by facilitating access to this information by developing information retrieval systems (IRS), IR has evolved with the emergence of the Web and more recently social networks (SN). Nowadays, RS is the most used way to communicate, share knowledge and content on the Web. With this social dimension that enriches the content of resources on the Web, users find themselves with new information needs. Hence the emergence of the Social RI (SIR), a recent theme that aims to take into account information specific to the SN. The main objective of our work is to integrate the different signals from social networks into the IR process in order to improve research and meet the needs of the user. Indeed, in this report, we present two models of social IR in which we propose two approaches that exploit the social signals of the social network Twitter, the first one is interested in indexing Tweets and the second in the use of these Tweets to extand the original query.

Keywords : Information retrieval, social information retrieval, social networks, social signals, Twitter.

Table des matières

Table des tableaux et des figures

Introduction générale

1. Contexte et problématique

La RI est une activité dont le but est de sélectionner un ensemble de documents à un utilisateur en fonction de son besoin en informations exprimé à l'aide d'une requête. Le défi est de pouvoir, parmi le volume important de documents disponibles, trouver ceux qui correspondent au mieux à l'attente de l'utilisateur.

L'essor du web a remis la RI face à de nouveaux défis d'accès à l'information, il s'agit cette fois de retrouver une information pertinente dans un espace diversifié et de taille considérable. Ces difficultés ont donné naissance à une nouvelle discipline appelée recherche d'Information sociale. La recherche d'information sociale permet de combiner entre une pertinence textuelle classique et une pertinence sociale issue des réactions des utilisateurs sur les ressources du Web, La motivation derrière l'exploitation de ces contenus, en particulier les signaux, sur la performance des systèmes de recherche d'information (SRI) est d'essayer de tirer profit de ces traces provenant des actions collectives des utilisateurs pour améliorer la RI par rapport à un besoin en information. Les principales problématiques liées à cette discipline consistent d'abord, à identifier les ressources sociales issues des réseaux sociaux pouvant répondre aux exigences de l'utilisateur et comment les exploiter pour améliorer le processus de la RI.

Nos travaux se situent dans la recherche d'information sociale, plus précisément, à l'exploitation des signaux sociaux de *Twitter* pour améliorer la RI.

2. Contribution

Notre contribution dans le cadre de la RI sociale, s'articule autour de deux points principaux :

1. Proposition d'une approche basée sur l'indexation des *Tweets* en utilisant leurs données sociales. En effet, nous proposons dans le cadre de l'intégration des signaux sociaux dans le processus de la RI, une approche permettant d'indexer les *Tweets* en utilisant ses signaux associés. Plus précisément, nous proposons une formule de pondération des termes de chaque *Tweet* qui prend en compte, en plus de leurs pertinence textuelles (TFIDF), leurs pertinences sociales exprimées par différents signaux sociaux (nombre de j'aimes, de retweets, de commentaires, etc.).
2. Proposition d'une approche basée sur la reformulation de la requête. Notre contribution dans cette nouvelle approche, consiste à utiliser *Twitter* comme collection externe afin d'étendre la requête initiale.

   Nous proposons de sélectionner, à partir de la requête initiale, un ensemble de *Tweets* jugés pertinents tout en utilisant la première approche pour l'indexation de ces derniers. Ensuite, les utiliser pour étendre la requête originale, et enfin, nous exploitons la requête

reformulée pour répondre aux besoins de l'utilisateur en utilisant la recherche traditionnelle dans une autre collection de documents.

3. Organisation du mémoire

Ce mémoire est organisé selon le plan suivant :

- **le chapitre 1** introduit les concepts de base de la recherche d'information (RI). Nous commençons par quelques définitions, ensuite nous allons présenter le processus en U de la RI en détaillant ses principales étapes. Enfin, nous concluons par quelques modèles de la RI et par les mesures d'évaluation des systèmes de recherche d'information (SRI).

- **Le chapitre 2** présente la recherche d'information sociale. Nous décrivons d'abord l'information sociale dans le Web. Ensuite, la notion de la RI sociale sera définie en mettant en évidence ses concepts de base. Puis, nous présentons un aperçu sur les travaux de l'état de l'art consacrés à l'exploitation des informations sociales dans le processus de RI. Enfin, nous présentons les principales collections de tests qui sont utilisées pour évaluer les SRI sociale.

- **Le chapitre 3** présente notre contribution qui concerne l'intégration et l'exploitation des signaux sociaux de *Twitter* au sein du processus de la RI. Nous présentons les deux approches proposées qui s'intéressent respectivement à l'indexation des *Tweets* et à la reformulation de la requête.

Chapitre 1 : Concepts de base de la RI

Introduction

Avec l'augmentation rapide du volume documentaire stocké sous format numérique, il est devenu très difficile de trouver une information ou un document qui répond à un besoin utilisateur, pour remédier à ce problème la recherche d'Information est apparue. Elle représente une discipline relativement ancienne qui date des années 1950, son objectif consiste à relier les documents d'une collection avec un besoin en information formulé par un utilisateur. Il incombe aux modèles de RI d'estimer une mesure statistique qui permet de quantifier la pertinence d'un document par rapport à ce besoin.

Ce chapitre a pour but de présenter le domaine de la RI. Nous allons commencer par présenter les concepts de base de la RI en décrivant son processus (processus en U). Ensuite, nous passeront en vue quelques modèles de la RI. Enfin, nous allons présenter un aperçu des mesures d'évaluation des systèmes de recherche d'information (SRI).

I. Définitions

- **Définition de la recherche d'information (RI)** : Gerard Salton (Salton G. , 1968) a défini la recherche d'information comme suit *« la recherche d'information est un domaine qui concerne la structure, l'analyse, l'organisation, le stockage, la recherche et la récupération des informations ».*

"Information retrieval is a field concerned with the structure, analysis, organization, storage, searching, and retrieval of information."

- **Définition du système de recherche d'information (SRI)** : un système de recherche d'information a été défini par (Kowalski, 1998) comme *« un système capable de stocker, de récupérer et de gérer des informations. Ces informations peuvent être composées de texte (y compris des données numériques et de date), des images, de l'audio, de la vidéo et d'autres objets multimédias »*.

"An Information Retrieval System is a system that is capable of storage, retrieval, and maintenance of information. Information in this context can be composed of text (including numeric and date data), images, audio, video and other multi-media objects"

II. Processus en U de la RI

Le processus de RI qui permet, à partir d'une requête, d'ordonnancer les documents est appelé "processus en U". Il est décomposé en trois principales étapes : l'indexation, le requêtage (recherche) et l'appariement.

La figure 1 illustre le processus en U de la RI :

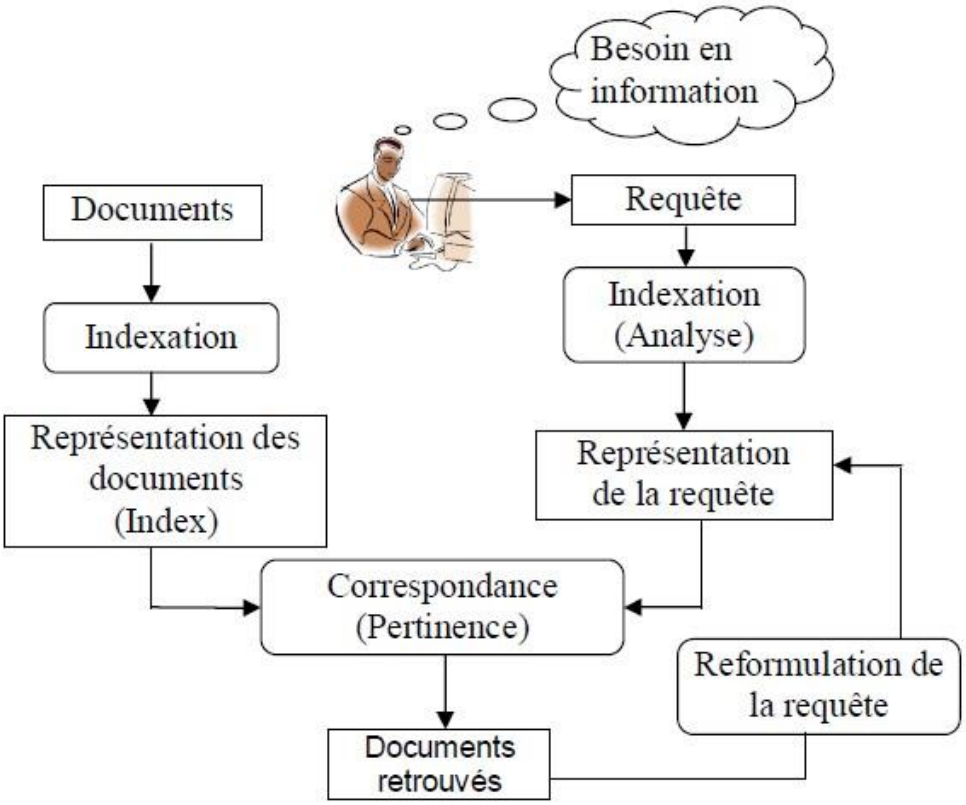

Figure 1 : Processus en U de la RI (Hamache, 2013)

Documents et collection de documents (corpus) : un document est un support physique de l'information, qui peut être du texte, une page web, une image, une séquence vidéo, etc. L'ensemble des documents manipulés par les SRI se nomme collection (corpus) de documents.

Requête : la requête constitue l'expression du besoin en information de l'utilisateur. Elle représente l'interface entre le SRI et l'utilisateur. Elle est souvent formulée à travers une liste de mots-clés.

II.1. Indexation

L'objectif de l'indexation est de permettre de retrouver rapidement les documents contenant les termes (mots-clés) de la requête. Elle consiste donc, à associer à chaque document une liste de mots clés (descripteurs), susceptible de représenter au mieux le contenu sémantique des documents. Il existe trois méthodes d'indexation :

II.1.a. Méthodes d'indexation

II.1.a.1. Indexation manuelle

L'indexation manuelle est réalisée par un expert qui identifie les termes les plus représentatifs du document. Cette méthode assure une meilleure qualité de résultats. Cependant, elle demande un travail manuel qui est non seulement très difficile mais très long à réaliser par les indexeurs.

II.1.a.2. Indexation semi-automatique

L'indexation semi-automatique se divise en deux parties, une partie automatique permettant d'extraire une liste de descripteurs, et une deuxième partie manuelle réalisée par un spécialiste du domaine dont la tâche est de sélectionner des termes significatifs parmi les descripteurs retournés auparavant. Généralement le spécialiste utilise un vocabulaire contrôlé sous forme de thésaurus ou de base terminologique.

II.1.a.3. Indexation automatique

L'indexation automatique est adoptée par la majorité des SRI en raison de son coût réduit, elle repose sur une démarche algorithmique qui traite chaque terme selon un processus défini, basé essentiellement sur une approche statistique.

II.1.b. Processus d'indexation

Le processus d'indexation comprend une série de traitements automatisés qui sont appliqués sur le document. Nous le détaillons dans ce qui suit.

II.1.b.1. Extraction des mots

Dite aussi tokenization du texte, est une première étape importante dans ce processus, elle est aussi bien appliquée au texte du document qu'à la requête. L'objectif de cette étape est d'identifier les mots (tokens) en reconnaissant tout ce qui représente des séparateurs, des caractères spéciaux, des chiffres, les ponctuations, etc.

II.1.b.2. Élimination des mots vides

Afin d'extraire les termes significatifs de ceux qui ne le sont pas dans un document, nous procédons par élimination des mots vides (les conjonctions, les prépositions, certains adverbes, etc..) car ils ne traitent pas le sujet du document. En supprimant ces termes, la taille de l'index sera réduite et son efficacité sera améliorée.

II.1.b.3. Normalisation

La normalisation consiste à représenter les différentes variantes d'un terme par un format unique appelé lemme ou racine qui porte un concept commun, par exemple les mots : économie, économiquement et économiste, peuvent être représentés par économie. Grâce à cette lemmatisation, les documents contenant différentes formes d'un même mot auront les mêmes chances d'être restitués. Par conséquent, elle réduit la taille de l'index et améliore le rappel (Badache, 2016).

II.1.b.4. Pondération des mots

Elle consiste à affecter à chaque terme $\boldsymbol{t_i}$ d'un document $\boldsymbol{d_j}$ un poids $\boldsymbol{w_{ij}}$, ce poids exprime le degré de représentativité du terme dans le document ce qui reflète l'importance du terme (Harrathi, 2010). Les termes sont généralement pondérés en utilisant des schémas de pondération tels que $TFIDF$ qui s'exprime par le produit de deux fonctions TF et IDF comme suit :

$$TFIDF = TF \times IDF$$

- $\boldsymbol{TF_{ij}}$ (Term Frequency) : appelée pondération locale, elle reflète l'importance locale du terme dans le document, en d'autres termes, elle représente la fréquence d'apparition du terme $\boldsymbol{t_i}$ dans le document $\boldsymbol{d_j}$. Cette formule de pondération peut être calculée différemment comme suit :

$$\mathrm{T}F_{ij} = \mathrm{f}\ (t_i,\ d_j)$$

$$\mathrm{T}F_{ij} = 1 + \log\ (\mathrm{f}\ (t_i,\ d_j))$$

$$TF_{ij} = \frac{f(ti, dj)}{\sum_k f(ti, dj)}$$

  Où f $(t_i,\ d_j)$ est la fréquence du terme $\boldsymbol{t_i}$ dans le document $\boldsymbol{d_j}$.

- **IDF** (Inverse Document Frequency) : c'est une pondération globale, elle est exprimée en nombre total de documents dans la collection et en fonction du nombre de documents contenant le terme. En effet, un terme présent dans certains documents

du corpus (collection) est plus important qu'un terme présent dans tous les documents. Cette mesure peut être calculée différemment comme suit :

$$IDF_i = \log(\frac{N}{n_i})$$

$$IDF_i = \log(\frac{N - n_i}{N})$$

Où n_i est le nombre de document contenant le terme $\boldsymbol{t_i}$ et N est le nombre total de documents dans le corpus.

II.2. Requêtage

La recherche vise à sélectionner les documents pertinents qui couvrent les besoins d'information de l'utilisateur. En effet, cette étape s'intéresse à l'expression des besoins de l'utilisateur, souvent à travers une liste de mots-clés représentant la requête. Divers types de langages d'interrogation ont été proposé en RI pour formuler une requête. Cette requête peut donc être exprimée en langage naturel ou quasi naturel, dans un format structuré, appelé aussi interrogation en langage booléen (exemple : "recherche d'information ET les réseaux sociaux"), ou à partir d'une interface graphique.

II.3. Appariement

La fonction d'appariement consiste à comparer la représentation de la requête avec les représentations des documents afin de mesurer la valeur de pertinence entre eux. Ce processus se base sur une fonction de correspondance (similarité) notée $RSV(q, d)$ (*Retrieval Status Value*) entre une requête « q » et un document « d ». Le résultat de cette comparaison se traduit par un score qui détermine la probabilité de pertinence (degré de similarité ou degré de ressemblance) du document vis-à-vis de la requête, qui permettra ensuite au SRI de renvoyer à l'utilisateurs les documents susceptibles d'être pertinents. Cette fonction de pertinence diffère d'un modèle de RI à un autre, c'est d'ailleurs ce qui les différencie entre eux. Parmi ces modèles on distingue trois principales catégories : modèle booléen, modèle vectoriel et le modèle probabiliste que nous détaillons dans ce qui suit.

II.4. Reformulation de la requête

L'utilisateur est parfois confronté à une situation difficile dans laquelle il est incapable de trouver les mots précis pour formuler sa requête et par conséquent les résultats retournés ne sont pas tous intéressants. Afin de palier à ce problème, les SRI doivent intégrer une autre fonctionnalité qui consiste à reformuler la requête. Le principe de ce processus est de modifier la requête initiale de l'utilisateur en rajoutant des termes significatifs ou/et la réestimation de leurs

poids. Si les termes ajoutés proviennent des documents de la collection, on parle donc de « *relevance feedback* » ou réinjection de pertinence. Par contre, s'ils sont issus d'une ressource conceptuelle externe (ontologie [1], thésaurus [2] ou dictionnaire [3]), on parle, dans ce cas, de reformulation directe de requête. Nous détaillons tout cela dans la suite de ce chapitre.

III. Modèles de RI

III.1. Modèle booléen

Le modèle booléen est historiquement le premier modèle de la recherche d'information (Salton, 1971). Il se base sur la théorie des ensembles et l'algèbre de Boole (Hamache, 2013). Dans ce modèle, la requête est représentée sous forme d'une expression logique dont les termes d'indexation sont reliés par des opérateurs booléens : OU (∨), ET (∧) et NON (¬) ; et le document est représenté par un ensemble de termes d'indexation (Harrathi, 2010).

La fonction de pertinence $RSV(q, d)$ entre une requête et un document est la vérification de l'implication logique $d \rightarrow q$. Autrement dit, un document est dit pertinent seulement si la fonction $RSV(q, d) = 1$ (appariement exacte) sinon il est considéré comme non pertinent.

Malgré la large utilisation de ce modèle et sa simplicité de mise en œuvre, il présente un certain nombre d'inconvénients dus à l'estimation binaire de la pertinence :

- Il ne permet pas de retourner des documents plus ou moins pertinents qui pourraient être utiles à l'utilisateur.
- Les documents retournés à l'utilisateur ne sont pas ordonnés selon leurs degrés de pertinence.
- Puisqu'il s'agit d'un appariement exact, il est donc difficile à l'utilisateur de bien formuler sa requête pour répondre à ses besoins.

Pour remédier à ces limites, des extensions ont été apporté à ce modèle en considérant la mesure de pertinence comme étant non-binaire. Parmi ces extensions on trouve le modèle booléen étendu (Salton, Fox, & Wu, 1983), et le modèle basé sur les ensembles flous.

[1] Ontologie : c'est une représentation formelle d'un domaine. C'est une conceptualisation, dans le sens où elle fournit un vocabulaire formalisé de concepts et de leurs relations.

[2] Thésaurus : il constitue un dictionnaire hiérarchisé des vocabulaires contrôlés dont les termes sont organisés dans une hiérarchie de concepts liés par des relations sémantiques.

[3] Dictionnaire : représente une structure constituant le langage d'indexation, il contient l'ensemble des termes reconnus par le SRI.

III.2. Modèle vectoriel

Proposé par Salton et ses collègues pour remédier aux faiblesses du modèle booléen, il repose sur la représentation algébrique des documents et de la requête (Soulier, 2014), ce qui permet d'avoir un appariement plus précis avec différents degrés de similarité.

Dans ce modèle, les documents et la requête sont représentés par des vecteurs dans l'espace engendré par tous les termes de l'indexation. Chaque document est représenté par un vecteur :

$$D_j = (\omega_{1j}, \omega_{2j}, \dots, \omega_{Mj})$$

De même pour la requête :

$$Q_i = (\omega_{1i}, \omega_{2i}, \dots, \omega_{Mi})$$

Avec ω correspond au poids d'un terme dans le document D_j ou dans la requête Q_i. le degré de similarité entre un document et une requête dans ce modèle, est déterminé par une mesure qui exprime le rapprochement entre le vecteur D_j et Q_i. Les principales mesures de similarité qui sont utilisées sont :

- ***Le produit scalaire***

$$\text{RSV}(Q_i, D_j) = \sum_{k=1}^{M} \omega_{ki} \cdot \omega_{kj}$$

- ***La mesure de Jaccard***

$$\text{RSV}(Q_i, D_j) = \frac{\sum_{k=1}^{M} \omega_{ki} \cdot \omega_{kj}}{\sum_{k=1}^{M} \omega_{ki}^2 + \sum_{k=1}^{M} \omega_{kj}^2 - \sum_{k=1}^{M} \omega_{ki} \cdot \omega_{kj}}$$

- ***La mesure de Cosinus***

$$\text{RSV}(Q_i, D_j) = \frac{\sum_{k=1}^{M} \omega_{ki} \cdot \omega_{kj}}{\sqrt{\sum_{k=1}^{M} \omega_{ki}^2} \cdot \sqrt{\sum_{k=1}^{M} \omega_{kj}^2}}$$

- ***La mesure de Dice***

$$\text{RSV}(Q_i, D_j) = \frac{2 \times \sum_{k=1}^{M} \omega_{ki} \cdot \omega_{kj}}{\sum_{k=1}^{M} \omega_{ki}^2 + \sum_{k=1}^{M} \omega_{kj}^2}$$

En plus de sa facilité de mise en œuvre, l'avantage de ce modèle vectoriel c'est qu'il permet une correspondance partielle ou approximative entre le document et la requête. il permet aussi de trier et de classer les résultats de la recherche selon le degré de similarité, ce qui offre une meilleure qualité de résultats pour l'utilisateur. Cependant, la représentation vectorielle considère chaque terme séparément alors qu'on peut avoir des termes qui sont en relation sémantique entre eux, ce qui induit à une baisse de précision du système.

III.3. Modèle probabiliste

III.3.a. Modèle probabiliste de base

Le modèle probabiliste est conçu par Maron and Kuhns en 1960, fondé sur la théorie des probabilités. Son principe de base est d'estimer la probabilité de pertinence d'un document vis-à-vis d'une requête. Autrement dit, il vise à retrouver les documents qui ont en même temps une forte probabilité d'être pertinents, et une faible probabilité d'être non pertinents (Abbassi, 2013). Un document est alors retourné si la probabilité qu'il soit pertinent notée $P\ (R/D)$ par rapport à une requête Q est supérieur à la probabilité qu'il ne soit pas pertinent $P\ (\bar{R}/D)$ (Harrathi, 2010). Le score de pertinence entre un document D et une requête Q est estimé comme suit :

$$\text{RSV}(Q_i, D) = \frac{P\ (R/D)}{P\ (\bar{R}/D)}$$

Après simplification en utilisant la formule de Bayes on obtient :

$$\text{RSV}(Q_i, D) = \frac{P\ (D/R)}{P\ (D/\bar{R})}$$

Avec $P\ (D/R)$ est la probabilité que le document D appartienne à l'ensemble R des documents pertinents et $P\ (D/\bar{R})$ est la probabilité que D appartienne à l'ensemble des documents non pertinents.

Différentes méthodes sont utilisées pour estimer ces deux probabilités, nous nous intéressant particulièrement au modèle d'indépendance binaire BIR (*Binary Indepance Retrieval*), où on considère que la variable document $D = (t_1 = x_1, t_2 = x_2, \dots, t_n = x_n)$ est représentée par un évènement : $(x_i = 1)$ si x_i est présent et par $(x_i = 0)$ si x_i est absent. Et en supposant que

les termes d'indexation sont indépendants, les deux probabilités précédentes peuvent être exprimées comme suit (Garrouch, 2017) :

$$P\ (D/R) = \prod_{i=1}^{i=n} P\ (t_i = x_i/R)$$

$$P\ \ (D/\bar{R}) = \prod_{i=1}^{i=n} P\ (t_i = x_i/\bar{R})$$

Avec t_i est le terme qui décrit le document D , et x_i est sa valeur (0, 1). la distribution des termes suit une loi de Bernoulli, $P\ (D/R)$ et $P\ \ (D/\bar{R})$ peuvent alors s'écrire comme suit :

$$P\ (D/R) = \prod_{i=1}^{i=n} P\ (t_i = 1/R)^{x_i} \times P\ (t_i = 0/R)^{1-x_i}$$

$$P\ (D/\bar{R}) = \prod_{i=1}^{i=n} P\ (t_i = 1/\bar{R})^{x_i} \times P\ (t_i = 0/\bar{R})^{1-x_i}$$

En prenant $p_i = P\ (t_i = 1/R)$ et $q_i = P\ (t_i = 1/\bar{R})$, et après simplification, la fonction $\mathrm{RSV}(Q_i, D)$ peut s'écrire comme suit :

$$\mathrm{RSV}(Q_i, D) = \sum_{i, x_i = 1} \log\left(\frac{p_i\ (1 - q_i)}{q_i\ (1 - p_i)}\right)$$

L'avantage de ce modèle probabiliste de base est qu'il permet la modélisation explicite de la notion de pertinence. Mais en revanche, il représente des inconvénients du fait qu'il ne prend pas en compte la fréquence des termes et la relation entre eux (indépendance entre les termes) (Boughanem, Modèles probabilistes pour la recherche d'information, 2015).

III.3.b. Modèle de langue

Une autre façon de modéliser les documents sous forme probabiliste réside dans l'utilisation du modèle de langue. Le principe de ce modèle consiste à créer un modèle de langue pour chaque document noté M_d , puis calculer la probabilité qu'une requête Q puisse être générée par le modèle de langue de document qu'on note $P\left(\frac{Q}{M_d}\right)$. Cette probabilité peut être mesurée comme suit (Abbassi, 2013) :

$$RSV\left(\frac{d}{Q}\right) = P\left(\frac{Q}{M_d}\right) = \prod_{i=1}^{i=n} P\ (t_i/D)$$

Où $t_i \in Q \; et \;\; P\left(\frac{t_i}{D}\right)$ peut être estimé en se basant sur l'estimation maximale de vraisemblance, ainsi on aura :

$$P(t_i/D) = \frac{TF(t_i, D)}{\sum_t TF(t, D)}$$

Le problème qui se pose dans ce type d'estimation est que si un terme de la requête est absent dans le document, la mesure de probabilité sera nulle ($P\left(\frac{Q}{M_d}\right) = 0$). Pour remédier à ce problème, des techniques de lissage peuvent être utilisées. Le principe de ces techniques consiste à assigner des valeurs non nulles aux termes non présents dans le document. En utilisant par exemple le lissage par interpolation de Jelinek-Mercer (Jelinek, 1980), la formule de probabilité précédente peut être réécrite comme suit :

$$RSV\left(\frac{d}{Q}\right) = \prod_{i=1}^{i=n} \left(\lambda_i \times P(t_i/D)\right) + \left((1 - \lambda_i) \times P\,(t_i)\right)$$

Avec λ_i est la probabilité que le terme à la position i soit important et $(1 - \lambda_i)$ est la probabilité que le terme ne soit pas important.

IV. Reformulation de la requête

Comme nous l'avons vu précédemment, la reformulation de la requête a pour principe de modifier la requête initiale formulée par l'utilisateur, par ajout des termes significatifs et/ou réestimation de leurs poids. Nous distinguons principalement, trois types de reformulation, une reformulation directe, une reformulation indirecte par injection de pertinence et une reformulation par pseudo relevance feedback.

IV.1. La reformulation directe

Elle consiste à ajouter de nouveaux termes à la requête initiale, soit grâce aux liens de cooccurrence entre les termes ou bien en se basant sur une ontologie capable, à partir d'une requête initiale, de l'enrichir avec des termes dérivant des relations sémantiques telles que la synonymie, la spécialisation/généralisation et la composition.

IV.2. La reformulation par injection de pertinence (*relevance feedback*)

Cette reformulation permet une modification de la requête initiale sur la base des jugements de pertinence de l'utilisateur, sur les documents restitués par le système. Son principe fondamental est d'utiliser la requête initiale pour amorcer la recherche d'information, puis

exploiter itérativement les jugements de pertinence de l'utilisateur, afin d'ajuster la requête par expansion ou repondération.

L'un des algorithmes d'expansion de requêtes les plus utilisés dans le domaine de la RI, est celui développé par Rocchio (J.J.Rocchio, 1971) basé sur le modèle vectoriel, où la nouvelle requête est générée en tenant en compte la distribution des termes dans les documents pertinents et non pertinents selon la formule suivante :

$$Q_1 = \alpha Q_0 + \beta \frac{1}{|D_p|} \sum_{D_j \epsilon\, D_p} D_j - \delta \frac{1}{|D_{np}|} \sum_{D_j \epsilon\, D_{np}} D_j$$

Avec :

Q_1 Est la nouvelle requête.

Q_0 Est la requête initiale.

D_p et D_{np} Représentent respectivement les documents jugés pertinents et non pertinents.

$|D_p|$ et $|D_{np}|$ Représentent respectivement le nombre de documents jugés pertinents et non pertinents.

α, β, δ sont des paramètres de la reformulation.

IV.3. La reformulation par pseudo relevance feedback

C'est une variante de la technique précédente qui est indépendante de l'utilisateur. Dans ce cas, le système suppose que les documents qui se trouvent en tête de la liste des documents retournés pour la requête initiale, sont pertinents et appliquent la réinjection de pertinence pour générer une nouvelle requête. Cette technique est utilisée lorsqu'aucun jugement utilisateur n'est disponible.

V. Évaluation des SRI

L'évaluation des SRI est une étape importante, elle permet de vérifier l'efficacité des modèles mis en œuvre pour l'identification des documents pertinents et à satisfaire le besoin en information de l'utilisateur. L'évaluation sert donc, à comparer les résultats retournés par le système (pertinence système) avec les attentes de l'utilisateur (pertinence utilisateur). Plus les réponses du système correspondent à celles que l'utilisateur espère, mieux est le système.

V.1. Collection de test

Une collection (ou corpus) de test constitue le moyen d'évaluation des SRI. Elle est généralement composée d'une collection de documents, une collection de requêtes et des

jugements de pertinence associés à ces requêtes. La création d'une collection de test n'est pas une tâche aisée : il faut trouver des documents qui doivent être libres de droits ou distribuables à la communauté sous certaines conditions, associer des requêtes "intéressantes", mais également et surtout, effectuer des jugements de pertinence pour chaque requête, ce qui se fait par un groupe d'évaluateurs qui examinent le contenu de chaque document et de juger s'il est pertinent par rapport à la requête.

Nous citons certaines compagnes d'évaluation qui sont les plus connues :

- La collection Cranfield est la première collection de tests permettant de mesurer quantitativement avec précision l'efficacité de la recherche d'informations. Elle est aujourd'hui trop petite pour des expériences pilotes les plus élémentaires. Elle contient 1398 résumés d'articles de revues aérodynamiques, un ensemble de 225 requêtes et des jugements de pertinence exhaustifs de toutes les paires (requête, document).
- La compagne TREC (*Text REtrieval Conference*) a été mise en place par NIST (Institut National des Standards et Technologies) en 1992. Ces collections comprenaient initialement 1,89 millions de documents et de jugement de pertinence pour 450 besoins d'informations. Elle regroupe à ce jour un large panel de tâches, telles que la recherche ad-hoc, ou également les tâches de recherche dans les microblogs ou celles orientées pour les systèmes de questions-réponses.
- La compagne CLEF (*Cross Language Evaluation Forum*) initialement construite pour des évaluations dans des langues européennes ainsi que la recherche d'information multi-langues. Elle est aujourd'hui, devenue un forum pour de nombreuses évaluations spécialisées comme la recherche d'information dans le domaine de la chimie ou des maths. Et en plus de proposer des tâches de recherche sur des documents, cette campagne fournit également des collections d'images associées à des annotations.

V.2. Mesures d'évaluation

Les mesures d'évaluation permettent d'estimer quantitativement l'efficacité d'un système. L'objectif est d'identifier, pour chaque requête la capacite du système à retourner des documents pertinents. Nous citons ci-dessous les mesures d'évaluations les plus utilisées.

- ***Précision et Rappel (Precision and Recall) :*** le rappel mesure la capacité du système à retrouver tous les documents pertinents répondant à une requête. Il est calculé comme étant un rapport du nombre de documents pertinents restitués (*SP*) sur le nombre total

de documents pertinents (P). Tandis que la précision mesure la capacité du système à rejeter tous les documents non pertinents à une requête donnée, par le rapport du nombre de documents pertinents restitués par le système (SP) sur le nombre total de documents restitués (R).

$$Rapport = \frac{SP}{P}$$

$$Précision = \frac{SP}{R}$$

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 ne se produit pas dans un système réel car le taux de précision et de rappel sont antagonistes comme le montre la figure 2 ci-dessous. En effet, Lorsque la précision augmente, le rappel diminue et inversement. Ainsi, pour mesurer les performances d'un système il faut utiliser les deux mesures conjointement.

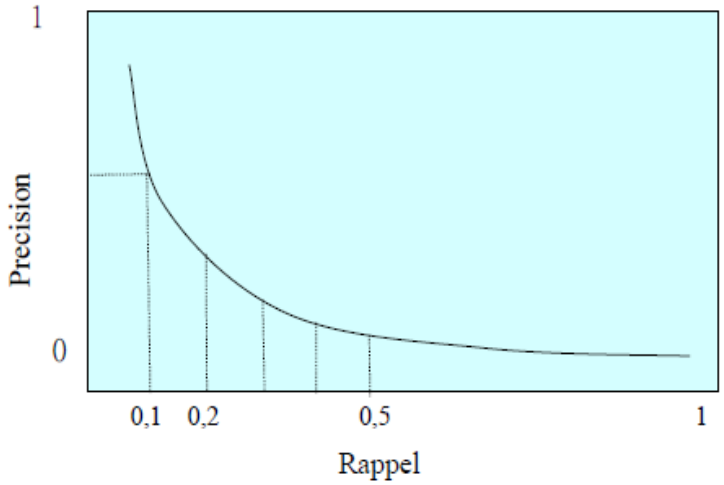

Figure 2 : Forme générale de la courbe de précision-rappel d'un système de RI (Boughanem, 2015)

- ***F-mesure*** *:* la F-mesure (F) est une moyenne harmonique pondérée de la précision et du rappel. Sa fonction est donnée comme suit :

$$F = \frac{1}{\alpha \cdot \frac{1}{précision} + (1-\alpha) \cdot \frac{1}{rappel}}$$

Avec $\alpha \in [0, 1]$

- ***Mean Average Precision (MAP) :*** cette mesure calcule la moyenne des valeurs de précision moyennes non interpolées sur l'ensemble des documents pertinents. La formule suivante donne la méthode de calcul de la MAP :

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

Avec Q est l'ensemble des requêtes, $|Q|$ est le nombre total de requêtes et AP_q est la moyenne des précisions à chaque rang de document pertinent pour une requête q. Elle est donnée par la formule suivante :

$$AP_q = \frac{1}{R} \sum_{i=1}^{N} p(i) \times R(i)$$

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 est le nombre de documents pertinents pour la requête q et N le nombre de documents restitués par le système.

Conclusion

La recherche d'information abordée dans ce chapitre est une RI classique. Elle ne s'intéresse qu'aux données textuelles. Cependant, avec l'émergence du web, plus précisément l'apparition des réseaux sociaux, l'enjeux de la recherche est devenu beaucoup plus important. Elle ne doit pas se limiter qu'aux données textuelles, mais il faut prendre certains autres critères en considération, comme les signaux sociaux (j'aime, partage, commentaire …), la temporalité, etc. Ce qui est détaillé dans le prochain chapitre qui porte sur la recherche d'information sociale (RIS).

Chapitre 2 : La recherche d'information sociale

Introduction

La recherche d'information (RI) est un domaine qui consiste à définir des modèles et des processus dont le but de retourner, à partir d'un corpus de documents indexés, ceux dont le contenu correspond le mieux au besoin en information exprimé par un utilisateur. Initialement développée pour des corpus de documents textuels, la RI a évolué avec l'émergence du Web et plus récemment des réseaux sociaux (RS). De nos jours, les RS représentent le moyen le plus utilisé pour la communication, le partage de connaissance et de contenus sur le Web. Avec cette dimension sociale qui vient enrichir les contenus des ressources sur le Web, les utilisateurs se retrouvent avec de nouveaux besoins en information. La RI classique ne semble pas adaptée à cette dimension, impliquant les utilisateurs et leurs interactions au sein des réseaux sociaux, d'où l'émergence de la RI Sociale (RIS), une thématique récente qui a pour objectif de prendre en compte les informations spécifiques aux RS.

III. Information sociale dans le WEB

III.1. Les médias sociaux

De nos jours, les réseaux sociaux sont le cœur du web 2.0, ils peuvent être définis comme étant un espace dans lequel les internautes peuvent interagir avec le contenu du web (publications, partages, annotation, collaboration, commentaires …) (Chahrazed, 2011), ils représentent aussi un moyen de communication et d'échange efficace entre les utilisateurs (amis, followers, abonnés…). En effet, les internautes sont passés de simples consommateurs dans le web 1.0, à des producteurs de l'information. L'ensemble de ces informations généré par l'utilisateur est appelé UGC (User Generated Content).

Parmi les différents réseaux sociaux qui existent nous listons ci-dessous les plus populaires :

- **Facebook[1] :** est le plus célèbre des réseaux sociaux, créé par Mark Zuckerberg en 2004, initialement destiné uniquement aux étudiants de l'université d'Harvard dans laquelle il faisait partie, puis il s'est étendu au monde entier en 2006. Il compte désormais plus de 2,32 milliards d'utilisateurs actifs chaque mois**.** Facebook permet aux utilisateurs, une fois inscrit, de créer un profil, publier des informations (photos, vidéos, liens, statuts …), de communiquer avec leurs familles, leurs amis et collègues grâce à la messagerie instantanée, réagir aux informations publiées par les autres par des j'aime, commentaires

[1] https://www.facebook.com/

ou partages, ils peuvent même créer des groupes ou des pages visant à faire connaitre des institutions ou des entreprises.

- ***Twitter*[2] :** Créé en 2006 par Jack Dorsey, portant le slogan « Que faites-vous en ce moment ? », l'idée de départ est simplement de partager le quotidien de ses utilisateurs avec un certain nombre de personnes appelés abonnés, par la suite, il est devenu un moyen de partage d'information en temps réel (*Tweets*). En fin 2018 Twitter compte 321 millions d'utilisateurs actifs mensuels qui publient 500 millions de *Tweets* chaque jour.
- **Instagram[3] :** Fondé par Kevin Systrom en octobre 2010, il permet de partager des photos et des vidéos avec son réseaux d'amis, et réagir par « j'aime » ou commentaire sur les publications des autres. Instagram a connu un succès instantané, en 2018 il compte plus d'un milliard de membres et plus 70 millions de photos partagées quotidiennement.
- **LinkedIn[4] :** a été fondé en décembre 2002 et lancé en mai 2003 par Reid Hoffman et Allen Blue, c'est un réseau social professionnel en ligne, les utilisateurs exposent dans leurs pages : leurs carrières professionnelles, leurs vies sociales et leurs loisirs, ainsi leur permettant de créer des liens entre autres professionnels. En 2018 il compte plus de 500 millions d'utilisateurs dans le monde, ainsi que plus de 150 secteurs d'activité dans 200 pays.
- **Pinterest[5] :** Créé en 2010 par Paul Sciarra, Evan Sharp et Ben Silbermann. Il compte plus de 200 millions d'utilisateurs actifs par mois, pour un chiffre d'affaires estimé à 550 millions de dollars en 2017. Il permet à ses utilisateurs de partager leurs centres d'intérêt et passions à travers des albums de photographies glanées sur Internet. Sachant que le nom du site est un mot-valise des mots anglais "pin" et "interest" signifiant respectivement "épingler" et "intérêt".
- **Wikis :** Un wiki est une application web qui permet la création, la modification et l'illustration collaboratives de pages à l'intérieur d'un site web. Les pages peuvent être visualisées dans deux modes différents : le mode lecture, qui est le mode par défaut, et le mode écriture, qui présente la page sous une forme qui permet de la modifier. Le plus célèbre des wikis est incontestablement Wikipédia, une encyclopédie libre en ligne qui est l'un des sites web les plus consulté du monde.
- **Blogs :** Le Blog, nommé par contraction des mots Web Log (carnet de bord web en anglais), est un site web dans lequel un ou plusieurs auteurs publient au fil du temps des articles (aussi appelés post ou billets), organisés en catégories et affichés dans l'ordre

[2] https://twitter.com/
[3] https://www.instagram.com/
[4] https://www.linkedin.com/

chronologique inverse. Les visiteurs du blog peuvent ensuite commenter le contenu des articles.

III.2. Contenus générés par l'utilisateur

Avec le Web actuel et l'évolution des réseaux sociaux, les utilisateurs sont amenés à interagir avec le contenu du Web et de contribuer à son développement. Ils sont passés de simples consommateurs à de producteurs de l'informations qu'on appelle Contenus Générés par l'Utilisateur (UGC).

I.1.a. Définition

Ce terme « contenus générés par l'utilisateur » ou « *User Generated Content* » en anglais, devient populaire pendant l'année 2005, il désigne tout type de contenu publiés par les utilisateurs sur des plateformes web (les réseaux sociaux, les wikis, les blogs ...), tel que des images, des vidéos, du texte et du son, ainsi que d'autres types de contenu y compris la fourniture de métadonnées supplémentaires, pour les ressources en ligne telles que des descriptions, ou des termes créés par un ensemble d'utilisateurs afin d'enrichir une ressource par des tags, ou encore un commentaire, un avis.

I.1.b. Les signaux sociaux

Les signaux sociaux représentent l'un des types les plus populaires des UGCs. Ils peuvent être définis comme étants les interactions, les émotions, les relations et les comportements sociaux d'une personne avec une autre ou avec une ressource sur le Web à travers des fonctionnalités offertes par les réseaux sociaux (figure 3).

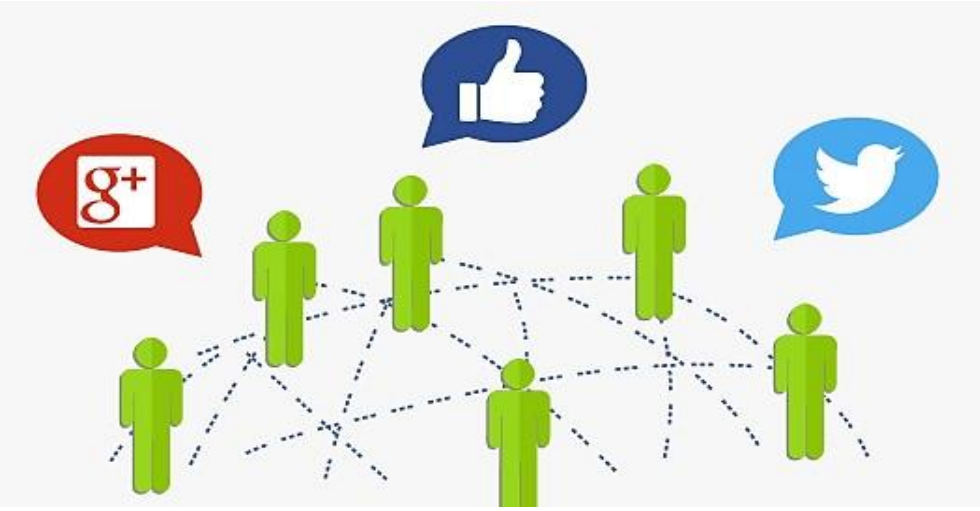

Figure 3 : l'interaction des signaux sociaux dans le web (Signaux sociaux, s.d.)

Il existe plusieurs types de signaux sociaux sur le web, chacun est associé à un réseau social particulier et dont la signification diffère d'un réseau à un autre. En effet, un partage sur Facebook et un *Retweet* sur *Twitter* n'ont pas la même signification ou la même influence sur le contenu du

web. Le tableau 1 ci-dessous regroupe les signaux sociaux les plus populaires sur les réseaux sociaux :

Type	**Example**	**Réseaux sociaux**
Votes	J'aime, +1	Facebook, Google+, LinkedIn, Instagram
Messages	*Tweet*, Publication	Facebook, Google+, LinkedIn, *Twitter*
Partages	Partage, *Retweet*	Facebook, Google+, LinkedIn, *Twitter*
Signets	Épingler	Pinterest
Commentaires	Commentaire, Répondre	Facebook, Google+, LinkedIn, Twitter
Relations	Abonnés, Amis	Facebook, *Twitter*, Instagram

Tableau 1 : Liste des différents types des signaux sociaux (Badache, 2016)

IV. Notion de la RI sociale

IV.1. Définitions

La recherche d'information sociale (RIS) est un domaine récent qui date des années 2000, après l'émergence du web 2.0 et la naissance des réseaux sociaux. Plusieurs définitions ont été proposé, Dion Goh et Schubert Foo (Goh & Foo, 2008) ont défini la RIS comme étant « *une famille de techniques de la recherche d'information qui aident les utilisateurs à répondre à leurs besoins en informations en exploitant l'expérience de recherche des autres utilisateurs* ».

« Social information retrieval refers to a family of information retrieval techniques that assist users in obtaining information to meet their information needs by harnessing other users' expert knowledge or search experience »

Sebastian Marius Kirsch (Kirsch, 2005) quant à lui, a défini la RIS comme « *l'incorporation d'informations sur les réseaux sociaux avec le processus de la recherche d'information* »,

« Social information retrieval is defined as the incorporation of information about social networks and relationships into the information retrieval process »

En 2009, James Lanagan (Lanagan, 2009) donne une autre définition qui décrit la recherche d'information sociale comme *« la création des annotations sur des contenus web existants est une forme d'interaction avec d'autres utilisateurs web conduit à ce qu'on appelle la recherche d'information sociale »*

« Creating annotations on existing web content is a form of interacting with other web users and leads to what is called social information retrieval »

IV.2. Concepts de RIS

Avec le web actuel et l'émergence des réseaux sociaux, les utilisateurs sont amenés à produire des informations qui sont rarement disponibles ailleurs dans les sites web ou dans des ressources bibliographiques. Cependant, les modèles classiques de la RI sont aveugles à ce contexte social qui entoure les utilisateurs et les ressources. Pour remédier à ce problème, les chercheurs ont été amené à créer un autre modèle de recherche qui se base sur les données sociales afin d'améliorer le processus de RI et mieux appréhender ses données, ce modèle est appelé la recherche d'information sociale (RIS) qui est illustré dans la figure suivante.

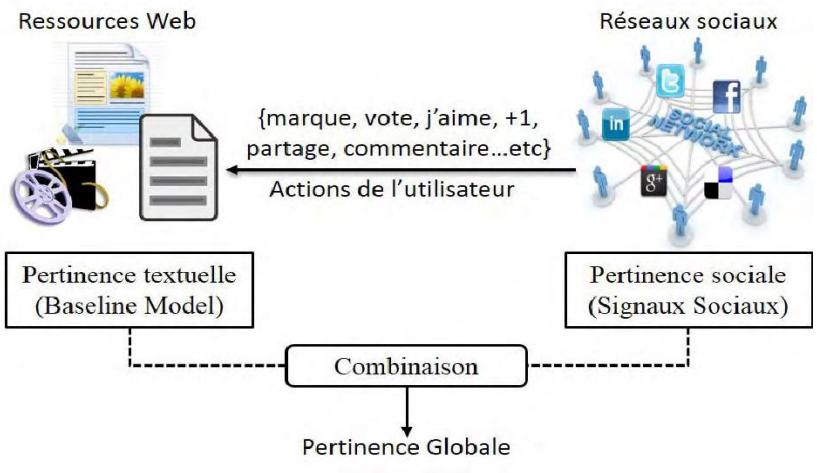

Figure 4 modèle de recherche d'information sociale (Badache, 2016)

En effet, les systèmes de RIS se distingue des autres SRI par l'utilisation des UGCs comme les signaux sociaux (j'aimes, commentaires, partages) pour satisfaire les motivations sociales derrière les besoins d'informations des utilisateurs. Les systèmes de RIS combine entre la pertinence textuelle classique et une pertinence sociale issus des réactions des utilisateurs sur les ressources du web. Il est donc nécessaire que la RIS puisse identifier les ressources sociales issues des réseaux sociaux, les exploiter pour améliorer la RI et les combiner afin de satisfaire les besoins en informations des utilisateurs (figure 5). Nous présentons dans ce qui suit les principaux

travaux de l'état de l'art consacrés à l'identification et à l'intégration des informations sociales au sein des modèle de RI à différents niveaux.

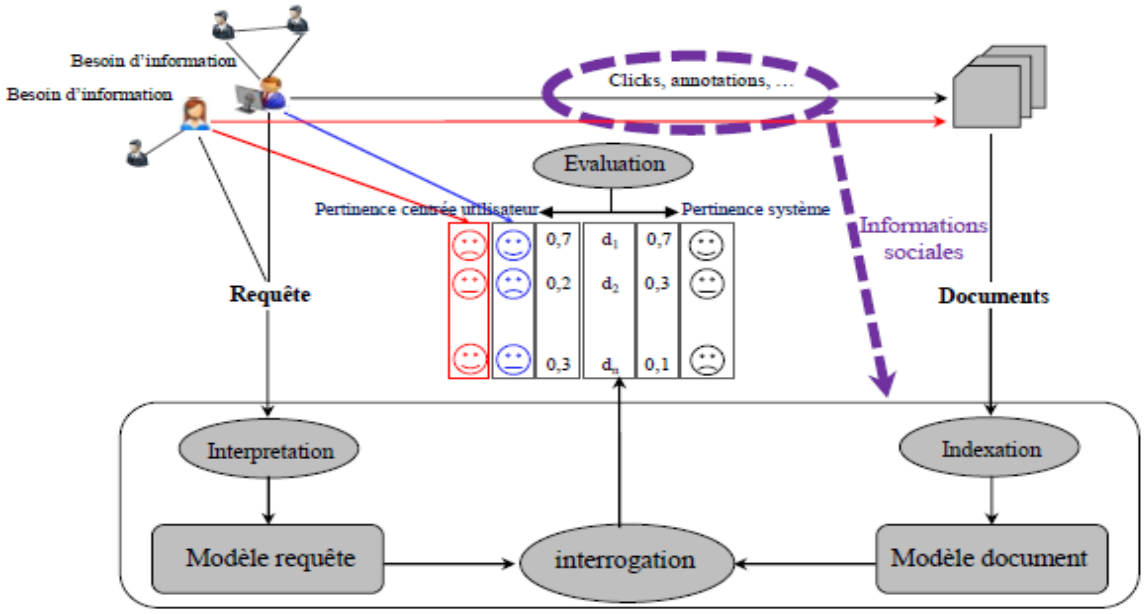

Figure 5 : Processus en U de la RIS (Chahrazed, 2011)

V. Les travaux de l'État de l'Art

V.1. Identification et exploitation des ressources sociales pour améliorer la RI

L'identification des ressources sociale a pour but de retrouver les informations sociales qui répondent aux exigences de l'utilisateur, c'est-à-dire, rechercher dans les différents espaces du web sociales les informations qui seront pertinentes ou qui sont susceptible de répondre aux exigences de l'utilisateur. En outre, les SRI sociale doivent analyser tout contenu généré par l'utilisateur en passant en revue les blogs, les microblogs, les réseaux sociaux (j'aimes, commentaires, publications, partages …)

Une fois l'information sociale identifiée et extraite, il faut trouver un moyen comment utiliser cette ressource pour améliorer le processus de la RI, nous présentons dans ce qui suit les travaux de l'état de l'art consacrés à l'intégration de ces informations sociales en distinguant ses différents niveaux d'amélioration.

V.1.a. Indexation sociale

Plusieurs travaux (Bischoff & al., 2008) (Dmitriev & al., 2006) ont démontré que l'ajout des annotations sociales tels que les tags au contenu du document améliorent efficacement les

résultats de la recherche, cette démarche est essentiellement utile si le document ne contient pas beaucoup de termes et que le processus d'indexation classique ne permet pas d'avoir une bonne performance de RI (Badache, 2016). Les travaux proposés dans ce contexte ont utilisé les annotations sociales de deux manières différentes :

- **Par l'ajout des données sociales aux contenu du document :** Certaines approches utilisent les métadonnées sociales pour enrichir le document. (Dmitriev & al., 2006), ont proposé d'utiliser annotations sociales dans la recherche intranet, pour que ces annotations puissent être intégrées au processus de la recherche il faudra les ajouter au contenu des pages. Cependant, cela ne prend pas en compte le fait que les annotations ont une sémantique différente et par conséquent, elles doivent être traitées comme des métadonnées, plutôt que du contenu. Pour cela, Dmitriev et ses collègues ont utilisé une base de données pour stocker les annotations explicites soumises par l'utilisateur. Cette base de données sert à afficher les annotations à l'utilisateur dans la barre d'outils et les résultats de la recherche. Périodiquement (une fois par jour), les annotations sont exportées dans un magasin d'annotations (annotations store) qui est un référentiel de documents au format spécial utilisé par leur système d'indexation. Le magasin d'annotations est combiné au magasin de contenu (content store) et au magasin de texte d'ancrage (Anchortext store) pour générer un nouvel index. Ceci est effectué en analysant séquentiellement ces trois magasins en mode batch et en utilisant un algorithme de fusion de tri basé sur disque pour construire l'index comme le montre la figure 6 ci-dessous.

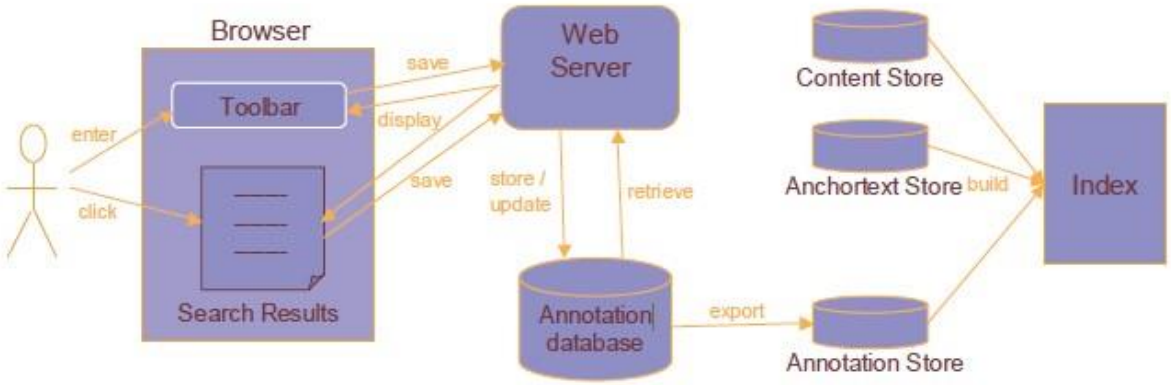

Figure 6 : Flux d'annotations dans le système (Dmitriev & al., 2006)

- **Par la représentation de documents personnalisés :** Compte tenu que chaque utilisateur a sa propre compréhension du contenu d'un document, il a donc tendance à employer un vocabulaire et des termes différents pour décrire, commenter et annoter ce document. Par conséquent, la solution est de créer un index pour chaque utilisateur afin de mieux répondre à son besoin en informations. Parmi les travaux proposés dans ce

contexte, nous citons le travail de Mohamed Reda Bouadjenek (Bouadjenek, 2013), qui a proposé d'utiliser une représentation personnalisée des documents sociaux (PSDR) et qui se fait en 3 phases :

- Chaque page web peut être représenté à l'aide d'une matrice $m \times n$ Users Tags matrix $M_{U,T}^{d}$ de m utilisateurs qui ont annoté la page web avec les n tags, et chaque entrée $w_{i,j}$ de la matrice représente le nombre de fois que l'utilisateur u_i a utilisé le terme t_j pout annoter la page web. Au lieu d'utiliser tous les tags des utilisateurs pour calculer le PSDR d'une page web, ils ont proposé de sélectionner uniquement les plus représentatifs afin de filtrer les utilisateurs non pertinents. Pour ce faire, ils ont utilisé une fonction de classement pour classer les utilisateurs de plus pertinents au moins pertinents et d'en sélectionner uniquement les top K, puis d'en sélectionner seulement les termes utilisés par le top K des utilisateurs, ainsi ils auront une nouvelle matrice Users-Tags réduite. La fonction de classement d'un utilisateur u selon un document d et une requête émise par un utilisateur u_q est donnée comme suit :

$$Rank_{u_q}^{d}(u) = \underbrace{\alpha \times \left(1 + log\left(|T_{u,d}|\right)\right) \times \log\left(\frac{|D|}{|D_u|}\right)}_{\text{Proximité de document}} + \underbrace{(1-\alpha) \times Sim(u, u_q)}_{\text{proximité de la requête}}$$

Avec $Sim(u, u_q)$ désigne la similarité entre l'utilisateur qui a annoté le document d et l'émetteur de la requête, α est un poids compris entre 0 et 1 qui permet de donner de l'importance soit au document soit à l'émetteur de la requête.

Une fois la matrice crée, ils ont procédé au calcul des poids $w_{i,j}$ associés à chaque entré de la matrice qui représentent la mesure dans laquelle l'utilisateur u_i estime que le terme t_j est associé au document d. Ils ont proposé d'utiliser une adaptation de la mesure de TF-IDF pour estimer ce poids. Par conséquent, ils ont défini le poids w_{t_i} du terme t_i dans un document d selon l'utilisateur u_i comme *user Term Frequency, inverse document Frequency (utf-idf),* qui est calculée comme suit :

$$w_{i,j} = utf - idf = log\left(1 + n_{u_i,t_i}^{d}\right) \times \log\left(\frac{|D_{ui}| + 1}{|D_{ui,ti}|}\right)$$

Où n_{u_i,t_i}^{d} est le nombre de fois que u_i a utilisé t_j pour annoter d et $|D_{ui}|$ (respectivement $|D_{ui,ti}|$) est le nombre de documents annotés par u_i (respectivement le nombre de documents annotés par u_i en utilisant le terme t_i.

- Chaque ligne de la matrice d'un document constitue la représentation sociale d'un utilisateur i . Toutefois, cette matrice est incomplète car elle contient de nombreuses valeurs manquantes. Le processus de factorisation permet ainsi

d'utiliser l'expérience et les feedbacks des autres utilisateurs afin de prédire les valeurs manquantes dans la matrice et ensuite calculer le PSDR de l'émetteur de requête en particulier.

- Enfin, Les PSDR doivent être appariés avec la requête pour quantifier leurs similitudes tout en considérant également le contenu textuel des documents. Par conséquent, ils ont proposé de calculer le score de classement personnalisé d'un document d qui correspond à une requête q émise par un utilisateur u en utilisant l'une des fonctions de classement suivantes :

  i) A Query Based Ranking Function (QBRF) :

  $$Rank(d,q,u) = \gamma \times \text{Sim}(\vec{q}, \overrightarrow{S_{d,u}}) + (1-\gamma) \times SES(\vec{d})$$

  ii) A Profile Based Ranking Function (PBRF) :

  $$Rank(d,q,u) = \gamma \times \text{Sim}(\overrightarrow{p_u}, \overrightarrow{S_{d,u}}) + (1-\gamma) \times SES(\vec{d})$$

  Où γ est un poids compris entre 0 et 1, $SES(\vec{d})$ est le Search Engine Score donné pour un document d, $\overrightarrow{S_{d,u}}$ est le PSDR de document d selon l'utilisateur u, et $\overrightarrow{p_u}$ est le profil de l'utilisateur u. Pour ce qui est de $\text{Sim}(\vec{q}, \overrightarrow{S_{d,u}})$ et de $\text{Sim}(\overrightarrow{p_u}, \overrightarrow{S_{d,u}})$ ils sont calculés comme suit en utilisant la mesure de cosinus :

$$\text{Sim}(\vec{q}, \overrightarrow{S_{d,u}}) = \frac{\vec{q} \cdot \overrightarrow{S_{d,u}}}{|\vec{q}| \times |\overrightarrow{S_{d,u}}|}$$

$$\text{Sim}(\overrightarrow{p_u}, \overrightarrow{S_{d,u}}) = \frac{\overrightarrow{p_u} \cdot \overrightarrow{S_{d,u}}}{|\overrightarrow{p_u}| \times |\overrightarrow{S_{d,u}}|}$$

V.1.b. Reformulation de la requête

La reformulation de la requête (Query Reformulation) consiste à réécrire la requête initiale soit en rajoutant d'autres termes significatifs si la requête initiale est ambigüe on parle donc d'expansion de requête (Query Expansion), soit en supprimant des termes ou des informations inutiles ce qu'on appelle raffinement de requête (Query Reduction).

Il n'y a pas de contributions de raffinement de requête en utilisant l'information sociale, mais tous les travaux existants se focalisent sur l'expansion de requête. Nous citons ci-dessous quelques travaux proposés dans ce contexte :

- Koolen et ses collègues (Koolen, Kazai, & Craswell, 2009) ont proposé une approche qui consiste à utiliser les données de Wikipédia comme collection externe pour étendre la requête et qui sera ensuite utilisée pour la recherche de livres. Leur travail consiste à

exploiter le chevauchement entre les requêtes de recherche et les titres des articles de Wikipédia pour sélectionner une page d'entrée pour l'extension de la requête. Bien que l'utilisation de plusieurs articles donne un plus grand vocabulaire de termes connexes parmi lesquels choisir, ils ont choisi d'utiliser un seul article spécifiquement sur le sujet de la requête en espérant conduire à des termes connexes plus proche des termes de la requête originale, et aboutissent ainsi à une précision accrue. De plus, ils pensent que les articles de Wikipédia sont souvent édités par plusieurs auteurs, qui ensemble, possèdent un vocabulaire plus étendu que chaque auteur individuellement.

Afin d'extraire des termes utiles pour l'expansion de requête de la page de requête associée dans Wikipédia, ils ont utilisé la formule TF-IDF dans le but de sélectionner les N termes qui décrivent le mieux le sujet d'une page donnée à partir des autres articles de Wikipédia. Le score TF-IDF d'un terme t est calculé comme suit :

$$tf.idf(t) = \frac{tf_d(t)}{|d|} \times \log\left(\frac{D}{df(t)}\right)$$

Où $tf_d(t)$ est la fréquence du terme t dans le document d, $|d|$ est la longueur du document d, D est le nombre total de documents dans la collection et $df(t)$ est le nombre de documents qui contiennent le terme t.

- Schenkel et al. (Schenkel & al., 2008) Proposent de sélectionner parmi l'ensemble des termes du contexte social de l'utilisateur un sous-ensemble de termes qui peuvent être employés pour étendre la requête de l'utilisateur. Ainsi, en plus de la dimension sociale sur laquelle les auteurs se basent pour construire un contexte social de l'utilisateur, ils proposent de prendre en compte une dimension sémantique pour sélectionner des termes. Cela se fait sur la base d'une similarité sémantique entre ces termes et ceux de la requête de l'utilisateur. Plus formellement, ils ont introduit la similarité de tags $tsim(t1, t2)$ pour une paire de tags (t1, t2) avec $0 \leq tsim(t1, t2) \leq 1$. Le score final d'un document d qui respecte un tag t par rapport à un utilisateur u et compte tenu de son expansion est défini comme suit :

$$s_u^*(d, t) = \max_{t' \in T} tsim(t, t') \times s_u(d, t')$$

Dans leur implémentation, la similitude entre deux tags est déterminée par la co-occurrence des tags dans la collection de documents entière en estimant des probabilités conditionnelles, ainsi :

$$tsim(t, t') = P[t'|t] = \frac{df(t \wedge t')}{df(t)}$$

Où $df(t \wedge t')$ est le nombre de documents qui sont tagués par les deux tags.

V.1.c. Reclassement des résultats

En RI, le classement des résultats consiste à définir une fonction de correspondance qui permet de calculer la similarité entre les documents et les requêtes. Nous distinguons deux

catégories pour le classement des résultats sociaux qui diffèrent dans la façon dont ils utilisent l'information sociale. La première catégorie utilise les informations sociales en ajoutant une pertinence sociale au processus de classement, tandis que la seconde l'utilise pour personnaliser les résultats de la recherche.

- **Classement basé sur la pertinence sociale :** cette première classe s'intéresse à la pertinence sociale d'un document en termes de popularité et réputation dans les réseaux sociaux. Diverses approches ont été proposé dans ce contexte comme SocialPageRank (Bao & al., 2007), FolkRank et the Adapted-Pagerank (Hotho & al., 2006), et SBRanK (Yanbe & al., 2007) qui sont tous une extension de l'algorithme de PageRank[6] (Brin & Page, 1998). La notion de base de l'algorithme Adapted-Pagerank est que si une ressource est taguée avec un nombre de tag important et par des utilisateurs importants, elle sera aussi importante. Hotho et ses collègues ont implémenté cet algorithme en deux étapes ; premièrement il transforme l'hypergraphe entre utilisateurs, tags et ressources en un graphe tripartite non orienté, pondéré et trié noté $G_F = (V, E)$ (avec V constitue l'ensemble disjoint utilisateur, tags et ressources et E représente les co-occurrences de tags et utilisateurs, utilisateurs et ressources, tags et ressources qui sont représentés par des arêtes non orientées et pondérées). Puis en deuxième lieu, ils appliquent une version de PageRank qui prends en compte les poids des arêtes sur ce graphe. Formellement, ils ont réparti le poids comme suit :

$$\vec{w} \leftarrow \alpha\vec{w} + \beta A\vec{w} + \gamma\vec{p}$$

Avec, A est la version row-stochastic de la matrice adjacente de GF, $\vec{p}$ est un vecteur de préférence (chaque ligne de la matrice est normalisée à 1 dans 1-norm), $\alpha, \beta, \gamma \in [0, 1]$ sont des constantes avec $\alpha + \beta + \gamma = 1$. La constante α est destinée à réguler la vitesse de convergence, tandis que la proportion entre β et γ contrôle l'influence du vecteur de préférence.

Ismail Badache (Badache, 2016) quant à lui, a proposé une approche qui consiste à estimer l'importance sociale d'une ressource en exploitant ses signaux sociaux associés, soit individuellement où chaque signal représente un facteur de pertinence, soit en regroupant ces signaux en fonction du type d'importance sous-jacent. Afin de prendre en compte ces facteurs sociaux dans l'évaluation de pertinence, il s'est appuyé sur un modèle de langue qui lui permet de combiner l'importance a priori de la ressource et sa pertinence vis-à-vis de la requête. La probabilité qu'une ressource D soit pertinente par rapport à une requête Q est estimée comme suit :

$$RSV(Q, D) = P(D|Q) = P(D) \cdot P(Q|D)$$

[6] PageRank : est un algorithme de Google qui mesure quantitativement la popularité d'une page web.

Avec $P(Q|D)$ qui représente la probabilité textuelle qui peut être calculé avec différent modèles tels que BM25 ou le modèle de langue, et $P(D)$ est une probabilité a priori indépendante de la requête qui est utile pour représenter et incorporer d'autres sources d'évidence dans le processus de RI comme les réactions sociales. Badache a proposé d'estimer $P(D_i)$ en considérant simplement le nombre de réactions sociales réalisées sur le document D_i. Ainsi, la formule générale pour calculer $P(D_i)$ est donnée comme suit :

$$P(D_i) = \prod_{r_j \in R} P\left(r_j \middle| D_i^r\right)$$

Où $P\left(r_j \middle| D_j^r\right)$ est lié à l'apparition de la réaction r_j dans le document D_i qui est calculé par le rapport entre le nombre de réaction r_j dans le document D_i sur le nombre total de réaction sur le même document.

- **Classement social personnalisé :** l'objectif de cette catégorie est d'améliorer la recherche en permettant de classer les documents différemment pour chaque utilisateur sous principe que les utilisateurs ont des profils et des besoins différents et donc ils ne devraient pas avoir les résultats classés de la même manière. Plusieurs travaux ont été proposé dans ce contexte (Bender & al., 2008), (Noll & Meinel, 2007), (Wang & Jin, 2010) et qui suivent la même idée qu'un score de classement d'un document d récupéré lorsqu'un utilisateur soumet une requête q est déterminé par :
    - Un processus de similarité entre les termes de la requête q et le document d pour générer un score de classement sans rapport avec l'utilisateur.
    - Un processus de mise en correspondance d'intérêts qui calcule la similitude entre un utilisateur u et le document d pour générer un score de classement lié à l'utilisateur.
    - Au final, une fusion est effectuée entre les deux classements précèdent pour avoir un classement final.

La technique de personnalisation de Noll et Meinel comprend deux étapes principales, à savoir, (i) la collecte et l'agrégation de données sur les utilisateurs et les documents, et (ii) la personnalisation de la recherche Web sur la base de ces données. L'agrégation des données collectées à partir des bookmarks et des tags se fait à travers deux profils :

- Un profil utilisateur : qui est représenté par une matrice tag-document M_d avec m tags et n documents et leurs n bookmarks.

$$M_d = \begin{bmatrix} c_{11} & \cdots & c_{1n} \\ \vdots & \ddots & \vdots \\ c_{m1} & \cdots & c_{mn} \end{bmatrix}, c_{ij} \in \{0,1\}$$

Un bookmark d'un document D_j est un vecteur b_j avec ses composants c_{ij} qui sont mis à 1 si le tag t_i est associé à d_j et à 0 sinon. Le profil utilisateur p_u est un vecteur avec m composants :

$$p_u := M_d \cdot \omega_d = \begin{bmatrix} c_1^* \\ \vdots \\ c_m^* \end{bmatrix}, c_i^* \in \mathbb{N}_0$$

Dans leur approche, ils ont défini le poids $\omega_d^T := 1^T = [1 \quad \dots \quad 1]$ avec n dimensions, attribuant ainsi une importance égale à tous les n documents, Et c_i^* représente le nombre total de tag t_i pour la collection de bookmarks de l'utilisateur.

- Un profil document : Contrairement aux profils utilisateur individuels, les profils de document sont un travail collaboratif. Chaque fois qu'un utilisateur crée ou modifie un bookmark d'un document Web, les informations sont partagées avec la communauté et le profil du document est mis à jour. Les métadonnées sur un document d peuvent être implémentées comme une matrice tag-user M_u avec m tags et n utilisateurs. Un bookmark d'un document d créé par l'utilisateur u_j est une colonne (vecteur) b_j dont les composants c_{ij} sont définis sur 1 si le tag t_i est associée à d par l'utilisateur u_j et à 0 sinon. Similaire aux profils utilisateur, le profil de document p_d est un vecteur avec m composantes générées par :

$$p_d := M_u \cdot \omega_u$$

Une fois les données collectées et transformées en profils utilisateur et profils de document, ils passent à la personnalisation pour mettre en correspondance les utilisateurs et les documents afin de réorganiser la liste des résultats de recherche, en suivant l'algorithme ci-dessous :

Algorithme : Personnalisation (utilisateurs, documents)

Require : le profil utilisateur et une séquence de liste de profils de documents.
Ensure : la liste personnalisée de documents pour l'utilisateur.
1 : for all d **in** documents **do**
2 : CALCULATE similarity (utilisateur, d)
3 : end for
tri des résultats de plus élevé au plus bas
5 : SORT documents **BY SIMILARITY**
6 : return documents

Où similarity (utilisateur, d) est défini comme : $similarity(u, d) := p_u^T \cdot \|p_d\|$, avec $\|p_d\|$ ramène simplement toutes les composantes non nulles de p_d à 1.

V.2. Exploitation de la temporalité des signaux sociaux pour améliorer la recherche

Les travaux cités précédemment ne s'intéressaient que sur les annotations sociales et leurs intégrations dans le processus de RI sans prendre en compte la notion de temps, autrement dit l'instant où l'action s'est produite ou la date de publication de la ressource. Peu de travaux ce sont intéressés à cette notion de temporalité, parmi eux nous citons :

- En plus des travaux d'Ismail Badache présentés précédemment (Badache, 2016), il a aussi proposé d'estimer l'importance sociale d'une ressource en exploitant le moment où l'interaction (signal) s'est produite ainsi que la date de publication de la ressource. Afin de prendre en compte cette importance dans l'évaluation de pertinence, il a repris le même modèle de langue basé sur les signaux sociaux qu'il a déjà présenté mais en prenant en compte l'aspect temporel.

    - **Prise en compte de la date du signal social :** dans ce niveau, Badache a proposé de compter les occurrences d'un signal en le pondérant (en le boostant) avec sa date d'apparition, soit Count_{ta} . La formule correspondante est la suivante :

$$\text{Count}_{ta}(t_{j,a_i}, D) = \sum_{j=1}^{k} f(t_{j,a_i}, D)$$

      Avec $t_{j,a} \epsilon T_{ai}$ et T_{ai} représente l'ensemble de k moments (date) à laquelle une action a_i s'est produite. La pondération de l'occurrence peut se faire de différentes manières, par exemple, prendre une fonction linéaire inversement proportionnelle à la date d'apparition :

$$f(t_{j,a_i}, D) = \frac{1}{t_{actuel} - t_{j;a_i}}$$

    - **Prise en compte de la date de publication de document** : qu'une vieille ressource a une plus grande chance d'avoir un grand nombre d'interactions par rapport à une ressource publiée récemment. Donc, pour résoudre ce problème, l'auteur a proposé de normaliser la distribution des signaux sociaux associés à une ressource par la date de publication de la ressource. La formule correspondante est la suivante :

$$\text{Count}_{tD}(a_i, D) = Count(a_i, D) \cdot A(D)$$

      Où :

$$A(D) = \exp\left(-\frac{\|t_{actuel} - t_D\|^2}{2\sigma^2}\right)$$

Avec $A(D)$ représente la fonction temporelle du document, estimée en utilisant le noyau Gaussien. Cette fonction calcule la distance temporelle entre la date actuelle t_{actuel} et la date de la ressource t_D. Et σ est un paramètre du noyau Gaussien.

- Inagaki et ses collègues (Inagaki & al., 2010) ont proposé d'exploiter les caractéristiques de clic en RI. Parmi ces critères, un facteur appelé Click Buzz, qui capte l'intérêt que suscite un document à travers le temps. Ils ont défini le Click Buzz comme une mesure pour déterminer si une page Web reçoit un niveau inhabituel d'intérêt des utilisateurs par rapport au passé. Le Click Buzz est basé sur le nombre de clics sur le document au cours d'un intervalle de temps donné. Cette méthode permet d'exploiter le feedback des utilisateurs pour améliorer la qualité des résultats de recherche en favorisant les URL qui ont un intérêt récent pour les utilisateurs.

VI. Évaluation de la RI Sociale

Comme nous l'avons vu dans le chapitre précédent, l'évaluation en RI se fait principalement à travers des collections de tests, souvent construites dans le cadre de campagnes d'évaluation. La RI sociale ne déroge pas à cette règle, avec la mise en place de la tâche Microblog dans la campagne d'évaluation TREC et la tâche de Social Book Search.

VI.1. La tâche TREC Microblog

Microblog Track examine les tâches de recherche et les méthodologies d'évaluation des comportements de recherche d'informations dans des environnements de microblogs tels que Twitter. Il a été introduit pour la première fois en 2011 et concernait une tâche de recherche ad hoc en temps réel (real-time Adhoc Search Task), dans laquelle l'utilisateur souhaite voir les informations les plus récentes mais pertinentes pour la requête. En 2011, la collection de texte *Tweets*2011 contenait 16 millions de *Tweets* (0,5 Go) exprimés dans diverses langues et publiés sur Twitter entre le 23 janvier 2011 et le 8 février 2011. L'ensemble de données est construit en utilisant l'API publique Twitter Stream qui fournit un échantillon représentatif de 1% du flux des *Tweets*. Ces dernières années, cette collection a évolué et le corpus a été fortement enrichi par de nouveaux *Tweets*. La collection actuelle, connue sous le nom *Tweets*2013, se compose de 243 millions de *Tweets* collectés à partir du flux public de Twitter entre le 1 Février et le 31 Mars 2013 (inclus).

VI.2. Social Book Search

La tâche de Social Book Search (SBS) étudie la recherche de livres dans des scénarios de recherche via une requête de l'utilisateur ou de recommandation. L'objectif est de modéliser et développer des techniques pour aider les utilisateurs dans les tâches de recherche de livres. La

tâche de Social Book Search se compose de trois Tracks : Interactive Track, suggestion Track et Mining Track.

- Interactive Track : est une tâche interactive orientée utilisateur qui examine les systèmes prenant en charge les utilisateurs à chacune des multiples étapes d'une tâche de recherche complexe. La Track offre aux participants une installation de recherche d'information interactive expérimentale complète et une nouvelle interface de recherche multi-étages passionnante pour étudier la manière dont les utilisateurs se déplacent au cours des étapes de recherche.
- Suggestion Track : une tâche orientée système qui consiste à suggérer des livres basés sur des requêtes de recherche riches combinant plusieurs signaux de pertinence thématique et contextuelle, ainsi que des profils d'utilisateurs et des jugements de pertinence réels.
- Mining Track : une piste NLP / Text Mining se concentrant sur la détection et la liaison de titres de livres dans des forums de discussion de livres en ligne, ainsi que sur la détection de recherches de livres dans des messages de forum pour la recommandation automatique de livres.

La collection INEX SBS se compose de 2.8 millions de documents. Chaque document décrit un livre d'Amazon, étendu avec des métadonnées sociales de LibraryThing. Chaque livre est un fichier XML représenté avec des champs comme ISBN, title, review, summary, rating and tag. La collection fournie 208 requêtes ainsi que leurs jugements de pertinence fournies par INEX.

Conclusion

Nous avons vu dans ce chapitre, que les réseaux sociaux sont des sources de précieuses informations appelées UGCs, qui peuvent être exploités pour améliorer de multiples services notamment la recherche d'information, ce qui a attiré de nombreux chercheurs à contribuer dans ce domaine pour donner naissance à la recherche d'information sociale. Nous avons ensuite, donné un aperçu sur les concepts de base de ce domaine. Puis, nous avons présenté les principaux travaux de l'État de l'Art basés sur les signaux sociaux et leurs intégrations dans la RI ainsi que l'aspect temporel qui a été pris en compte par quelques chercheurs.

Chapitre 3 : Exploitation des signaux sociaux de *Twitter* pour améliorer la RI

Introduction

Les premiers travaux en recherche d'information se sont basés essentiellement sur l'appariement entre les documents et la requête utilisateur, indépendamment de l'utilisateur et de son interaction avec ces documents. Cependant, avec l'émergence des réseaux sociaux et la disponibilité de différentes informations sociales, les travaux de RI ont commencé à s'intéresser davantage à l'exploitation de ces informations pour améliorer le processus de recherche, ce qui a donné naissance à la RI sociale. Comme nous l'avons vu dans le chapitre précédent, la majorité des approches proposées dans le cadre de la recherche sociale emploie différents facteurs de pertinence tels que le nombre de partages, de commentaires, de tags et la temporalité, etc.

Dans notre travail, nous nous intéressons à l'exploitation des signaux sociaux du réseau social *Twitter*. D'abord, nous présentons quelques généralités sur ce réseau social ainsi que son fonctionnement de base. Ensuite nous proposons deux modèles de recherche sociale, l'un s'intéresse à l'indexation des *Tweet* et l'autre à l'utilisation des *Tweets* comme ressource externe pour reformuler la requête de l'utilisateur. Nous détaillons ces deux approches dans ce chapitre.

I. Hypothèse

Notre objectif dans le cadre de nos travaux de recherche est d'étudier les différents facteurs de pertinence qui peuvent être utilisés pour améliorer les résultats de recherche. Nous nous focalisons dans ce travail sur l'exploitation des signaux sociaux de *Twitter* pour améliorer le processus de la RI. Notre hypothèse de recherche est abordée sous deux approches, dans la première nous supposons que *Twitter* peut être utilisé comme source d'informations directe, c'est-à-dire, pour rechercher une information ou une actualité on se réfère directement aux *Tweets* déjà publiés et on sélectionne ceux qui répondent à la requête. Dans la deuxième approche, nous utilisons *Twitter* non pas comme source d'information principale mais plutôt comme une ressource externe. En d'autres termes, la recherche ne se fait pas directement sur *Twitter* pour retourner des résultats, mais en revanche, son utilisation sert à mieux exprimer le besoin en information, considérant que dans les réseaux sociaux il y'a de forte chance que plusieurs utilisateurs expriment le même besoin en information en utilisant des termes différents.

II. Généralités sur le réseau social Twitter

Twitter est le plus populaire des réseaux sociaux, il a été créé en 2006 par Jack Dorsey, portant le slogan « Que faites-vous en ce moment ? ». Son objectif est de mettre en relation des utilisateurs grâce à des informations qu'ils postent eux-mêmes, appelées « *Tweets* » via le site web de Twitter, par téléphone mobile, ou encore via SMS. Ces informations sont de faible longueur (ne dépassent pas les 280 caractères) et qui peuvent être aussi bien des articles liés à l'actualité, aux centres d'intérêt, que des billets d'humeur.

II.1. Fonctionnement de *Twitter*

Twitter permet aux utilisateurs de créer un compte gratuitement et de le personnaliser en mettant une photo, une localisation, une biographie (pour permettre à *Twitter* de recommander des utilisateurs ayant les mêmes centres d'intérêts), etc. Une fois connecté, l'utilisateur peut s'exprimer à l'aide d'un court texte ne dépassant pas 280 caractères appelé *Tweet*, comme il peut aussi, rediffuser un autre *Tweet* dans lequel il peut y ajouter des informations complémentaires, il s'agit dans ce cas là d'un *Retweet*. Un utilisateur A peut suivre les flux envoyés par l'utilisateur B pour recevoir automatiquement toutes les publications (*Tweets*) de B, dans ce cas-là, A est appelé abonné (*Follower*) de B. Cette relation d'abonné peut être bidirectionnelle si B à son tour s'abonne à A pour recevoir tous les flux de ce dernier. Les *Tweets* partagés par un utilisateur, seront automatiquement visibles pour tous ses abonnés, et eux à leurs tours, peuvent aimer le *Tweet*, le commenter ou encore le *Retweet*er afin qu'il soit vu par tous leurs abonnés.

II.2. Les signaux sociaux de *Twitter*

Les différentes relations existantes entre les utilisateurs ainsi que leurs interactions avec les ressources sont considérées comme des signaux sociaux pouvant être utilisés comme facteur de pertinence dans la recherche d'information sociale. Nous listons dans le tableau 2 ci-dessous les divers signaux sociaux qui peuvent être extraits du réseaux social *Twitter*.

Les signaux sociaux	**Descriptions**
Nombre d'abonnés	Le nombre de personnes qui suivent l'utilisateur.
Nombre d'abonnements	Le nombre de personnes que l'utilisateur suit.
Nombre de j'aime d'un *Tweet*	Le nombre de personnes ayant aimé le *Tweet*
Nombre de commentaire	Le nombre de personnes ayant commenté le *Tweet*
Nombre de *Retweet*	Le nombre de personnes ayant partagé le *Tweet* avec ses abonnés

Tableau 2: les différents signaux sociaux de Twitter

III. Approches proposées

III.1. Architecture générale des approches proposées

Notre objectif dans le cadre de nos travaux de recherche, est de définir des approches permettant l'exploitation des signaux sociaux de Twitter comme facteurs de pertinence qui peuvent jouer un rôle pour améliorer la recherche d'information. Plus précisément, nous proposons une architecture de RI sociale structurée de la façon suivante (figure 7).

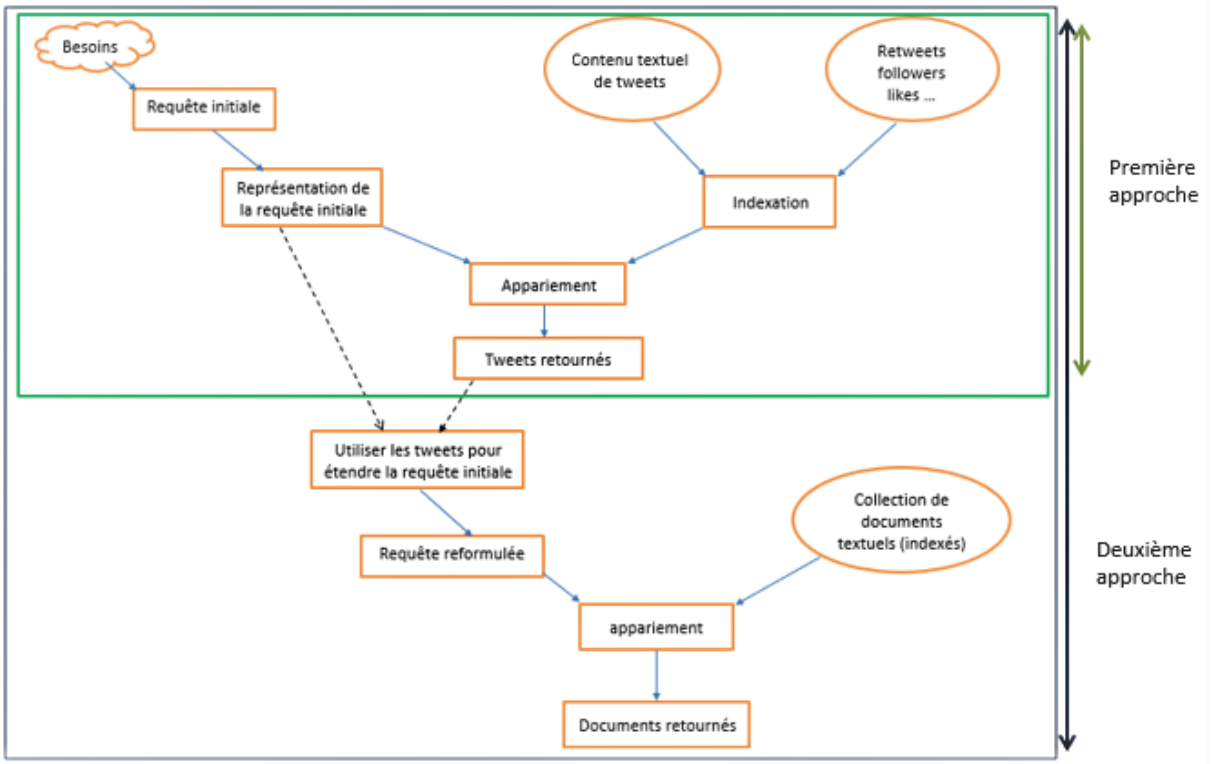

Figure 7 : Architecture des approches proposées

Dans l'architecture illustrée dans la figure 7, nous avons proposé deux approches exploitants les signaux sociaux de *Tweet*er afin d'améliorer les résultats de recherche. La première s'intéresse à l'ajout de ces signaux dans l'étape de l'indexation pour favoriser les *Tweet*s les plus populaires. La deuxième approche utilise les résultats retournés lors de l'appariement entre la requête et les *Tweet*s indexés pour reformuler la requête initiale afin de mieux exprimer le besoin en information de l'utilisateur en effectuant la recherche dans une collection externe.

III.2. Notations

L'information sociale que nous exploitons dans la première approche peut être représentée par un triplet $< U, M_s, T >$ où U, M_s, T sont des ensembles finis représentant respectivement *Utilisateurs, Métadonnées sociales, Tweets.* Idem pour la deuxième approche sauf que dans celle-ci, on rajoute un ensemble représentant les ressources, ainsi on aura un quadruplet $< U, M_s, T, R >$.

- ***Utilisateurs*** : cet ensemble représente les personnes qui interagissent avec le réseaux sociale *Twitter*. Ils peuvent publier des *Tweets*, aimer les *Tweets* des autres utilisateurs et les *Retweet*er, etc. Dans les travaux que nous présentons dans ce mémoire, on ne s'intéresse pas particulièrement à cet ensemble, mais nous l'avons introduit pour une éventuelle expansion de nos approches, à savoir la recherche sociale personnalisée d'information, qui s'intéresse principalement à cet ensemble utilisateurs.
- ***Métadonnées sociales :*** représente le nombre d'actions que les utilisateurs ont effectué sur les *Tweets* (nombre de j'aime, de commentaires, de *Retweet*es et le nombre d'abonnés de la personne qui a publié le *Tweet*), ainsi on aura : $M_s = \{N_{rt}, N_j, N_f, N_c\}$ où N_{rt}, N_j, N_f, N_c représentent respectivement le nombre de *Retweets*, le nombre de j'aimes, le nombre de followers (abonnés) et le nombre de commentaires.
- ***Tweets :*** est une collection de N *Tweets*, chaque *Tweet* partagé contient une information textuelle exprimée par un utilisateur U_i et un ensemble de métadonnées sociales (N_j, N_{rt}, N_f, N_c) relatives à ce *Tweet* et à la personne qui a partagé ce *Tweet*.
- ***Ressources :*** représente une collection de M documents, chaque document peut être du texte, une page web ou tout autre information numérique et qui est indépendante des *Tweets*.

III.3. Approche basée sur l'indexation des *Tweets*

Cette première approche consiste à modifier le processus de l'indexation des *Tweets* afin de prendre en compte, en plus de leurs contenus textuels, les métadonnées relatives à chaque *Tweet* dans le but de différencier entre eux par leurs popularités ou leurs réputations. Nous présentons dans ce qui suit, le processus d'indexation que nous avons proposé (figure 8).

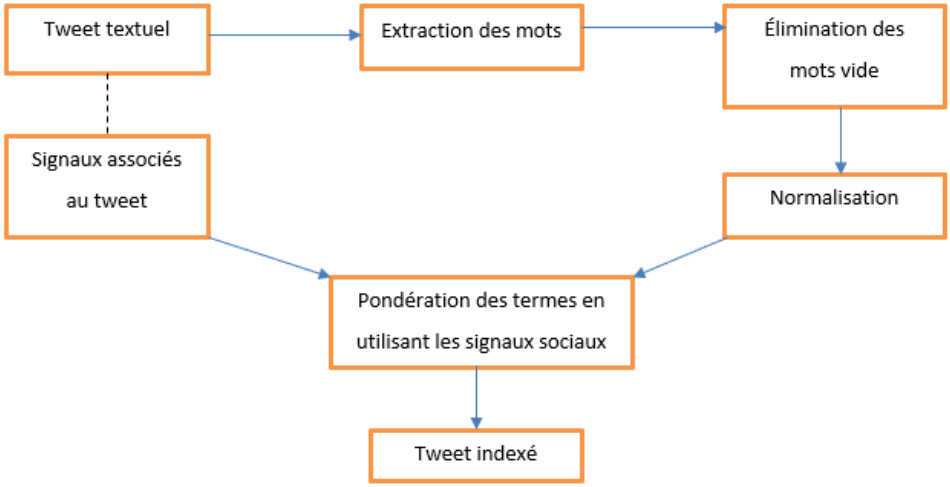

Figure 8 : Processus d'indexation des Tweets

- ***Extraction des mots :*** cela permet pour chaque *Tweet* d'extraire la liste des termes (tokens) qui le constitue en reconnaissant les espaces de séparation des mots, les chiffres, les ponctuations, etc.
- ***Élimination des mots vide :*** cette étape permet de réduire considérablement la taille de l'index en supprimant les mots non significatifs et non représentatifs du contenu informationnel des *Tweet*s. Ces mots sont souvent des prépositions, pronoms, certains adverbes, adjectifs, etc. Pour ce faire, on utilise un anti-dictionnaire (qui contient une liste de mots vides ou stoplist en anglais) et on vérifie si un terme dans le *Tweet* appartient à l'anti-dictionnaire il sera supprimé.
- ***Normalisation :*** elle consiste à représenter les différentes variantes d'un mot, par une forme unique appelée, lemme ou racine. Plusieurs stratégies de normalisation existent, dans notre approche nous utilisons l'algorithme de Porter (Porter, 1980) du fait que la majorité des *Tweet*s sont en anglais et que cet algorithme est le mieux adapté à cette langue.
- ***Pondérations des termes :*** Elle consiste à affecter à chaque terme d'un *Tweet* un poids $\boldsymbol{w_{ij}}$, ce poids exprime le degré de représentativité du terme dans le *Tweet* ce qui reflète l'importance du terme. Notre principale contribution dans cette première approche, consiste à modifier cette fonction de pondérations afin de prendre en compte les métadonnées relatives à chaque *Tweet*. Pour cela, nous proposons d'utiliser la fonction usuelle $TFIDF$, à laquelle on intègre le concept social de chaque *Tweet*. Cette fonction est donnée comme suit :

$$\omega = \alpha\ (TF * IDF) + (1 - \alpha) \cdot P(t)$$

Avec $P(t)$ représente un poids indépendant des termes et qui est relatif aux données sociales de chaque *Tweet*. Cette mesure peut être exprimé comme suit :

$$P(t) = \begin{cases} \left(\dfrac{\gamma N_{rt} + N_j + N_c)}{N_f}\right) \ si\ N_f > 1 \\ 0\ sinon \end{cases}$$

Où :

TF : fréquence du terme dans le *Tweet* (Term Frequency) $TF_{ij} = \frac{f(ti,Tj)}{\sum_k f(ti,Tj)}$.

IDF : fréquence du document inverse $IDF_i = \log(\frac{N}{n_i})$.

N_{rt} : Nombre de *Retweets* associé au *Tweet*.

N_j : Nombre de mentions « j'aime ».

N_c : Nombre de commentaires.

N_f : Nombre d'abonnés de la personne qui a partagé le *Tweet*.

α, et γ sont des poids, tel que $0.5 < \alpha < 1$ *et* $1 < \gamma < 3$. Le paramètre α est utilisé pour donner plus d'importance à la pondération textuelle, tandis que γ est utilisé pour favoriser les *Retweets* plus que les j'aimes et les commentaires. Les valeurs exactes de ces deux paramètres sont fixés de manière expérimentale.

Le principe de cette formule est de calculer le poids d'un terme en utilisant le $TFIDF$ standard (pertinence textuelle), en lui rajoutant une pertinence sociale indépendante du terme et qui est relative au *Tweet* auquel il appartient. Autrement dit, un terme i qui appartient à un *Tweet* j noté $t_{i,j}$, est plus pertinent si le *Tweet* j est considéré comme pertinent. Nous proposons de calculer cette pertinence de *Tweet* par un rapport entre le nombre de j'aimes, de commentaires et de *Retweets* regroupés, sur le nombre d'abonnés de la personne qui a partagé ce *Tweet*. En effet, nous supposons qu'un *Tweet* est pertinent si la majorité des abonnés de cet utilisateur ont réagi au *Tweet* soit par un j'aime, un commentaire ou encore un *Retweet*.

La structure algorithmique de cette approche peut être représentée comme suit :

ALGORITHME D'INDEXATION DES TWEETS :

```
BEGIN
    FOR EACH TWEET IN TWEETS :
        FOR EACH WORD IN TWEET :
            IF (WORD IN STOPLIST) THEN
                DELETE(WORD)
            ELSE
                PORTER_NORMALIZE(WORD)
                IF (N_f = 0) THEN
                    ω = 0.6 ( TF × IDF)
                ELSE
                    ω = 0.6 (TF × IDF) + (1 − 0.6)((γN_rt + N_j + N_c) / N_f )
                ENDIF
            ENDIF
        ENDFOR
    ENDFOR
END
```

III.4. Approche basée sur l'expansion de la requête

Dans cette deuxième approche, nous proposons d'utiliser Twitter comme une collection externe pour étendre la requête initiale, puis d'utiliser cette nouvelle requête afin de répondre aux besoins de l'utilisateur en utilisant la recherche classique.

La reformulation de la requête est un processus permettant la construction d'une nouvelle requête, plus à même de représenter les besoins en information de l'utilisateur. Elle est souvent opérée par ajout et/ou réévaluation des poids des termes de la requête initiale. Il existe principalement deux types de reformulation : reformulation direct et reformulation indirect (par injection de pertinence). Nous nous intéressant dans notre approche à l'injection de pertinence, plus particulièrement, à l'algorithme de Relevance Feedback de Rocchio (J.J.Rocchio, 1971). Il permet de construire une nouvelle requête, à partir de la requête initiale, et d'un ensemble de documents jugés pertinents et non pertinents. L'algorithme de Rocchio se base sur le modèle vectoriel, il permet de représenter la requête et les documents dans un espace vectoriel, puis il essaie de rapprocher au maximum la requête des documents jugés pertinents et de l'éloigner des documents jugés non pertinents, comme l'illustre la figure 9 ci-dessous.

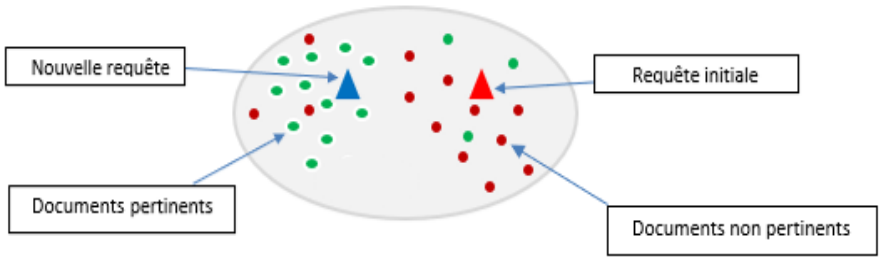

Figure 9 : Illustration de l'algorithme de Relevance Feedback de Rocchio

La formule de Rocchio pour l'injection de pertinence est donnée comme suit :

$$Q_1 = \alpha Q_0 + \beta \frac{1}{|D_p|} \sum_{D_j \epsilon\, D_p} D_j - \delta \frac{1}{|D_{np}|} \sum_{D_j \epsilon\, D_{np}} D_j$$

Avec :

Q_1 est la nouvelle requête.

Q_0 est la requête initiale.

D_p et D_{np} représente respectivement les documents jugés pertinents et non pertinent.

$|D_p|$ et $|D_{np}|$représente respectivement le nombre de documents jugés pertinents et non pertinent.

α, β, δ sont des paramètres de la reformulation, tels que $\alpha = 1$, $\beta = 0.5$, et $\delta = 0.25$ (ces valeurs ont été fixé dans d'autres travaux).

Une variante de cette technique a été proposé, nommée pseudo relevance feedback, qui ne s'intéresse qu'à rapprocher la requête des documents pertinents sans prendre en compte les documents non pertinents. Sa formule est exprimée comme suit :

$$Q_1 = \alpha Q_0 + \beta \frac{1}{|D_p|} \sum_{D_j \epsilon D_p} D_j$$

L'approche que nous proposons consiste d'abord à faire une recherche dans la collection de *Twitter* avec la requête de l'utilisateur tout en utilisant la première approche pour indexer les *Tweet* en utilisant leurs métadonnées sociales. Puis, d'en sélectionner les premiers *Tweets* les plus pertinents vis-à-vis de la requête nommés D_p (documents pertinents), que nous allons utiliser dans l'algorithme de Rocchio de pseudo relevance Feedback pour rapprocher la requête le plus possible vers les *Tweet*s pertinents. Cette formule est donnée comme suit :

$$Q_1 = \alpha Q_0 + \beta \frac{1}{|T_p|} \sum_{T_j \epsilon T_p} T_j$$

Avec :

Q_0 est la requête initiale.

T_p *Tweet*s jugés pertinents.

$|T_p|$ représente le nombre de *Tweet* pertinents (exemple $|T_p| = 10$).

α, β sont des paramètres de la reformulation ($\alpha = 1$, $\beta = 0.5$)

On aura comme sorti, une nouvelle requête Q_1 beaucoup plus représentative du besoin de l'utilisateur, qui est constituée des termes les plus pertinents extraits des *Tweets*. Cette nouvelle requête sera ensuite utilisée pour faire une recherche standard en utilisant la RI classique dans une autre collection hors que *Twitter*.

L'algorithme générale de cette approche peut être représenté comme suit :

ALGORITHME DE REFORMULATION DE LA REQUÊTE :

BEGIN
FOR EACH WORD IN QUERY :
$\omega = TF \times IDF$
ENDFOR
VICTOR_SPACE (QUERY)
FOR EACH TWEET IN TWEETS :
FOR EACH WORD IN TWEET :
IF (WORD IN STOPLIST) THEN
DELETE(WORD)
ELSE
PORTER_NORMALIZE(WORD)
IF ($N_f = 0$) THEN
$\omega = 0.6\ (TF \times IDF)$
ELSE
$\omega = 0.6\ (TF \times IDF) + (1 - 0.6)(\frac{\gamma N_{rt} + N_j + N_c)}{N_f})$
ENDIF
ENDIF
ENDFOR
VICTOR_SPACE (TWEET)
CALCULATE_RSV (QUERY, TWEET)
SORT_DESC_RSV (QUERY, TWEET)
ENDFOR
$T_p := TOP\_N\_RSV\ (QUERY, TWEET)$
FOR EACH TWEET IN T_p :
VICTOR_SPACE $(Query2)$ = VICTOR_SPACE $(query) + 0.5 \times \frac{1}{N}$(VICTOR_SPACE $(TWEET)$)
ENDFOR
FOR EACH DOCUMENT IN COLLECTION
FOR EACH WORD IN DOCUMENT
$\omega = TF \times IDF$
ENDFOR
VICTOR_SPACE (DOCUMENT)
CALCULATE_RSV (QUERY2, DOCUMENT)
ENDFOR
END

IV. Exemples et Tests

En premier lieu, nous proposons de tester l'indexation des *Tweet*s, sur quelques *Tweet*s extraits de Twitter sur le changement climatique, présentés dans le tableau 3 ci-dessous. Nous comparons entre l'indexation textuelle classique avec l'approche que nous proposons.

N° *Tweet*	Contenu	Nombre de j'aime	Nombre de commentaires	Nombre de *Retweet*	Nombre d'abonnés
1	Could reporters stop asking if political leaders "believe" in the danger of climate change and start asking if they understand it instead.	9200	827	4950	160820
2	Honestly climate changes scares the heck out of me and it makes me soo sad what we're loosing because of it.	2100	200	847	217000
3	There is no longer a worthy debate to be had about whether we have a problem. There is a worthy debate to be had about how we go about solving that problem.	166	40	1	6837
4	Climate change is real	552	3	552	7843

Tableau 3 : Liste des Tweets de test avec leurs caractéristiques sociales

- Indexation du *Tweet* N°01 :

Terme	fréquence du terme dans le Tweet	Nombre de Tweet contenant le terme	indexation classique (TF IDF)	indexation avec aspect sociale
reporter	1	1	0,050171666	0,079666487
stop	1	1	0,050171666	0,079666487
ask	2	1	0,100343332	0,109769486
political	1	1	0,050171666	0,079666487
leader	1	1	0,050171666	0,079666487
believe	1	1	0,050171666	0,079666487
danger	1	1	0,050171666	0,079666487
climate	1	3	0,010411561	0,055810424
change	1	3	0,010411561	0,055810424
start	1	1	0,050171666	0,079666487
understand	1	1	0,050171666	0,079666487
TOTAL	12			

Tableau 4 : indexation des termes du Tweet N°01

- Indexation du *Tweet* N°02

Terme	fréquence dans le document	nombre de documents contenant le terme	indexation classique TF IDF	indexation avec aspect sociale
honest	1	1	0,08600857	0,058967354
climate	1	3	0,017848391	0,018071247
changes	1	3	0,017848391	0,018071247
scares	1	1	0,08600857	0,058967354
makes	1	1	0,08600857	0,058967354
sad	1	1	0,08600857	0,058967354
loosing	1	1	0,08600857	0,058967354
TOTAL	7			

Tableau 5 : indexation des termes du Tweet N°02

- Indexation du *Tweet* N°03

Terme	fréquence dans le document	nombre de documents contenant le terme	indexation classique TF IDF	indexation avec aspect sociale
debate	2	1	0,240823997	0,156663478
problem	2	1	0,240823997	0,156663478
solve	1	1	0,120411998	0,084416279
TOTAL	5			

Tableau 6 : indexation des termes du Tweet N°03

- Indexation du *Tweet* N°04

Terme	fréquence dans le document	nombre de documents contenant le terme	indexation classique TF IDF	indexation avec aspect sociale
climate	1	3	0,041646246	0,109598228
change	1	3	0,041646246	0,109598228
real	1	1	0,200686664	0,205022479
TOTAL	3			

Tableau 7 : indexation des termes du Tweet N°04

- Indexation de la requête

Soit par exemple une requête utilisateur « climate change ». Nous indexons d'abord les termes de cette requête, puis nous calculons le score de similarité entre la requête et les *Tweets* indexés.

fréquence dans la requête	nombre de documents contenant le terme	indexation (TF x IDF)
1	3	0,062469368
1	3	0,062469368

Tableau 8 : Indexation de la requête

- Calcule de similarité *Tweets*/Requête en utilisant le produit scalaire

	sans prendre en compte l'aspect sociale	avec aspect social
score(Tweet1, requête)	0,001300807	0,006972884
score(Tweet2, requête)	0,002229955	0,002257799
score(Tweet3, requête)	0	0
score(Tweet4,requête)	0,005203229	0,013693064

Tableau 9 : Comparaison entre les scores (Tweets, requête)

Dans le tableau 9 ci-dessus, nous pouvons constater que l'utilisation des signaux sociaux de *Twitter* dans l'indexation a permis de donner un apport positif sur le classement des résultats retournés. En effet, les *Tweets* qui ont une proportion plus élevée entre le nombre d'abonnés et la somme des autres réactions ont été classé en premier, ce qui permet de les mettre en évidence. En d'autres termes, ils seront plus susceptibles de répondre aux besoins des utilisateurs.

Dans la deuxième partie des tests, nous nous intéressons à l'approche de la reformulation de la requête. Nous considérons les deux tweets les plus pertinents (Tweet 1 et 4) vis-à-vis de la requête en prenant en compte le facteur social. Ensuite, nous les utilisons dans la formule de Rocchio de pseudo relevance feedback présentée précédemment, pour étendre la requête initiale.

$$Q_1 = \alpha Q_0 + \beta \frac{1}{|T_p|} \sum_{T_j \epsilon\, T_p} T_j$$

On fixe $\alpha = 1, \beta = 0.5$

$|T_p|$ représente le nombre de tweets jugés pertinents, dans notre exemple $|T_p| = 2$

Terme	poids des terme de la requête initiale	poids des terme du Tweet N°01	poids des terme du Tweet N°04	poids des terme de la nouvelle requête
reporter	0	0,079666487	0	0,019916622
stop	0	0,079666487	0	0,019916622
ask	0	0,109769486	0	0,027442372
political	0	0,079666487	0	0,019916622
leader	0	0,079666487	0	0,019916622
believe	0	0,079666487	0	0,019916622
danger	0	0,079666487	0	0,019916622
climate	0,062469368	0,055810424	0,109598228	0,103821531
change	0,062469368	0,055810424	0,109598228	0,103821531
start	0	0,079666487	0	0,019916622
understand	0	0,079666487	0	0,019916622
real	0	0	0,205022479	0,05125562

Tableau 10 : application de l'algorithme de pseudo relevance feedback

Le tableau 10 ci-dessus, montre les poids des termes de la nouvelle requête en appliquant l'algorithme de pseudo relevance feedback. À partir de ces poids, nous pouvons prendre par exemple, les 3 premiers termes qui ont le poids le plus élevé pour les rajouter à la requête initiale. Nous aurons donc une nouvelle requête plus longue contenant les termes les plus pertinents extraits des tweets, par exemple :

Requête reformulée : {climate, change, real, ask, danger}

Considérons trois documents textuels. Nous prenons dans notre exemple, deux d'entre eux pour comparer entre le score de similarité de la requête initiale avec celui de la requête reformulée.

Document 1 : « Global climate change is real and measurable. Since the start of the 20th century, the global mean surface temperature of the Earth has increased by more than 0.7°C … »

Document 2 : « the climate has an impact on the mood, a sudden temperature variation can quickly change the attitude of a person »

- indexation du document 1

Terme	frequence du terme	nombre de documents contenant le terme	Indexation classique
global	2	1	0,086749319
climate	1	2	0,016008296
change	1	2	0,016008296
real	1	1	0,04337466
measure	1	1	0,04337466
mean	1	1	0,04337466
surface	1	1	0,04337466
temperature	1	2	0,016008296
earth	1	1	0,04337466
increase	1	1	0,04337466
TOTAL	11		

Tableau 11 : indexation du document 1

- indexation du document 2 :

Terme	fréquence du terme dans le document	nombre de documents contenant le terme	indexation classique TF IDF
climate	1	2	0,022011407
impact	1	1	0,059640157
mood	1	1	0,059640157
temperature	1	2	0,059640157
variation	1	1	0,059640157
change	1	2	0,022011407
attitude	1	1	0,059640157
person	1	1	0,059640157
TOTAL	8		

Tableau 12 : indexation du document 2

- Poids des termes de la nouvelle requête :

Terme	poids des termes
climate	0,103821531
changes	0,103821531
ask	0,027442372
real	0,05125562
danger	0,019916622

Tableau 13 : poids des termes de la requête reformulée

- Comparaison entre les résultats retournés par les deux requêtes :

	requête initale	requête étendue
score(document1, requête)	0,002000056	0,005547207
score(document2, requête)	0,002750077	0,004570516

Tableau 14 : comparaison entre résultats de recherche en utilisant la requête initiale et reformulée

Comme nous pouvons le constater d'après le tableau 14 ci-dessus, la requête reformulée a donné un apport positif sur les résultats retournés, contrairement aux résultats de la requête initiale.

Conclusion

Dans ce chapitre, nous avons présenté deux modèles de recherche d'information basés sur les signaux sociaux et leurs propriétés sociales. Ces signaux sont considérés comme étant une connaissance du document permettant de mesurer son intérêt et sa pertinence indépendamment de la requête. Afin de montrer la contribution de ces signaux sociaux dans la pertinence des documents, nous avons proposé deux approches qui exploitent les signaux sociaux de Twitter. La première s'intéresse à l'indexation des *Tweet*s en utilisant, en plus du contenu textuel, les signaux sociaux associés à chaque *Tweet* à savoir les *Retweet*s, les j'aimes, les commentaires et le nombre d'abonnés. Quant à la deuxième approche, celle-ci utilise les *Tweet*s dans la reformulation de la requête.

Conclusion générale

1. Conclusion générale

Les travaux présentés dans ce mémoire, rentrent dans le cadre de la recherche d'information, plus précisément dans la RI sociale. Les techniques traditionnelles de la RI se limitent à l'appariement entre documents et requête sans prendre, pour autant, l'aspect utilisateur ou sociale d'une ressource. Il devient donc, de plus en plus difficile de trouver une information pertinente, notamment avec le web actuel et l'émergence des réseaux sociaux. La problématique à laquelle nous nous sommes intéressé dans ce mémoire, réside à l'intégration et l'exploitation des informations sociales afin d'améliorer le processus de la recherche. Nous avons proposé dans ce mémoire deux approches :

1. Nous avons présenté une première approche qui exploite les signaux sociaux du réseau sociale *Twitter* dans le processus de RI, plus précisément à l'étape d'indexation. Nous avons proposé une formule de pondération des termes de *Tweets* qui prend en compte, en plus de contenus textuels des *Tweets,* l'aspect social de chacun d'eux. Cette formule est calculée en combinant entre une pertinence textuelle en utilisant la formule de pondération usuelle TFIDF, et une pertinence sociale indépendante des termes de Tweets exprimant leurs importances et qui est calculée à partir des différents signaux sociaux associés aux *Tweets*.
2. Dans la deuxième approche, nous nous sommes intéressés à la reformulation de la requête, plus particulièrement à son expansion en utilisant *Twitter* comme une collection externe. Le but de cette démarche est d'améliorer la représentation de la requête utilisateur afin de mieux exprimer son besoin et par conséquent améliorer les résultats de recherche. Nous nous sommes basés sur la première approche pour trouver les Tweets les plus pertinents vis-à-vis de la requête, ensuite les utiliser dans l'algorithme de Rocchio de pseudo relevance feedback afin d'étendre la requête initiale. Et par la suite, utiliser la requête reformulée pour faire une recherche dans une autre collection autre que *Twitter.*

   Cependant, une question importante que nous n'avons pas pu abordé, concerne l'expérimentation de nos deux approches en raison de l'indisponibilité de la collection TREC MICROBLOG et du temps nécessaire à ces expérimentation.

2. Perspectives

Dans les travaux futurs et en perspective, nous avons l'intention :

- En premier lieu, de faire des expérimentations sur nos deux approches et de les évaluer.
- En deuxième lieu, d'inclure l'aspect temporel dans les deux approches que nous avons proposé afin de favoriser les *Tweets* les plus récents et de ne pas les pénaliser, car les *Tweets* les plus anciens sont susceptible d'avoir plus de réactions que les nouveaux et par conséquent, ils seront mieux classés.
- Enfin, notre dernière perspective est de proposer une autre approche basée sur la recherche sociale personnalisée d'information, qui prend en compte chaque utilisateur séparément, sous prétexte que les utilisateurs ont des préférences différentes et donc ils peuvent avoir des résultats différents pour une même requête.

Bibliographie

Abbassi, M. (2013). *Un modèle de reformulation des requêtes pour la recherche d'information sur le Web.*

Badache, I. (2016). *Exploitation des Signaux Sociaux pour Am´eliorer la Recherche.* Toulouse: Université Toulouse 3 Paul Sabatier.

Bao, S., & al. (2007). *Optimizing Web Search Using Social Annotations.* Canada: IW3C2.

Bender, M., & al. (2008). *Exploiting Social Relations for Query Expansion and Result Ranking.* Saarbrücken: Max-Planck Institute for Informatics.

Bischoff, K., & al. (2008). *Can All Tags be Used for Search?* Hannover: L3S Research Center.

Bouadjenek, M. R. (2013). *Infrastructure and Algorithms for Information Retrieval Based On Social Network Analysis/Mining.* Versailles: University of Paris-Saclay.

Boughanem, M. (2015). *Evaluation des performances dans.* Toulouse: Université Toulouse 3 Paul Sabatier.

Boughanem, M. (2015). *Modèles probabilistes pour la recherche d'information.* Toulouse: Université Paul Sabatier.

Brin, S., & Page, L. (1998). *The anatomy of a large-scale hypertextual Web search engine.* California.

Carmel, D., & al. (2010). *Social bookmark weighting for search and recommendation.* Haifa: IBM Haifa Research Labs.

Chahrazed, B. (2011). *recherche d'inoformation.* Toulouse: université de Toulouse.

Damak, F. (2014). *Étude des facteurs de pertinence dans la recherche de microblogs.* Toulouse: Université Toulouse 3 Paul Sabatier.

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

Dmitriev, P. A., & al. (2006). *Using Annotations in Enterprise Search.* Edinburgh: IBM Almaden Research Center.

Garrouch, K. (2017). *Modèles de Recherche d'information basés sur les Réseaux Bayésiens et les Réseaux Possibilistes.* Tunisie: Université de SFAX.

Goh, D., & Foo, S. (2008). *Social Information Retrieval Systems : Emerging Technologies and Applications for Searching the Web Effectively.* New York: IGI Global.

Hamache, A. (2013). *Recherche d'Information : un modèle de langue combinant mots simples et mots composés.* Tizi-Ouzou: Université Mouloud Mammeri.

Harrathi, R. (2010). *Recherche d'information conceptuelle dans les documents.* Lyon: Institut Nationale des Sciences Appliquées de Lyon.

Hotho, A., & al. (2006). *Information Retrieval in Folksonomies: Search and Ranking.* Kassel: University of Kassel.

Inagaki, Y., & al. (2010). *Session Based Click Features for Recency Ranking.* Sunnyvale: Yahoo Labs.

J.J.Rocchio. (1971). *Relevance Feedback in information retrieval.* USA: The SMART retrieval system: experiments in automatic document processing.

Jelinek, F. (1980). *Interpolated estimation of markov source parameters from.* Pattern recognition in practice.

Kirsch, S. M. (2005). *Social Information Retrieval.* Bonn: Université Rheinische Friedrich-Wilhelms de Bonn.

Koolen, M., Kazai, G., & Craswell, N. (2009). *Wikipedia Pages as Entry Points for Book Search.* Barcelona,.

Kowalski, G. (1998). *INFORMATION STORAGE AND RETRIEVAL SYSTEMS.* Amherst: W. Brace Croft.

Lanagan, J. (2009). *Social Impact Retrieval: Measuring Author In Information Retrieval.* Dublin: Dublin City University School of Computing.

Noll, M. G., & Meinel, C. (2007). *Web Search Personalization via Social Bookmarking and Tagging.* Potsdam: Institut Hasso Plattner de l'Université de Potsdam.

Porter, M. (1980). *An algorithm for Suffix Stripping.*

Salton, G. (1968). *Automatic information organization and retrieval.*

Salton, G., Fox, E., & Wu, H. (1983). *Extended boolean information retrieval .* New York.

Schenkel, R., & al. (2008). *Efficient Top-k Querying over Social-Tagging Networks.* Singapore: SIGIR'08.

Signaux sociaux. (s.d.). Récupéré sur ryte.com: https://fr.ryte.com/wiki/Signaux_sociaux

Soulier, L. (2014). *Définition et évaluation de modèles de recherche d'information collaborative basés sur les compétences de domaine et les rôles des utilisateurs.* Toulouse: Université Toulouse 3 Paul Sabatier.

Wang, Q., & Jin, H. (2010). *Exploring Online Social Activities for Adaptive Search Personalization.* California: IBM Almaden Research Center.

Yanbe, Y., & al. (2007). *Towards Improving Web Search by Utilizing Social Bookmarks.* Kyoto: Department of Social Informatics, Kyoto University.