République Algérienne Démocratique Et Populaire

Université Mouloud MAMMERI de Tizi-Ouzou

Faculté de Génie Électrique et Informatique

Département d'Informatique

Mémoire de Fin d'Étude de MASTER

Thème:

« La recherche d'information temporelle dans les microblogs, cas: Twitter »

Présenté par :

****BENAMI Kenza***

****BEN OUALI Taoues***

Devant le jury composé de :

M^me^ AMIROUCHE. F …….. (Présidente)

M^me^ BELKACEMI .L ……...(Promotrice)

M^r^ SADI. S ……...(Examinateur)

2019/2020

Remerciements

En tout premier lieu, nous tenons à remercier Dieu le tout puissant et miséricordieux, qui nous a donné la force et la patience d'accomplir ce modeste travail.

Nous voudrons remercier notre directrice de mémoire Mme BELKACEMI Lila pour ses conseils qui ont contribué à alimenter notre réflexion.

Nous remercions également les membres du jury d'avoir accepté de juger notre travail.
Et également toute l'équipe pédagogique de l'université de Mouloud MAMMERI, et les intervenants professionnels et ainsi que les enseignants qui ont participé à notre formation.
Enfin, nous tenons à remercier toutes les personnes qui ont participé de près ou de loin à l'élaboration de ce modeste travail. Merci à vous tous … !

Dédicaces

Du plus profond de mon coeur je dédie ce travail à ceux qui me sont chers,

A mon très cher père BENAMI Nacer,

Au seul homme qui m'est précieux, la meilleure offre que le bon Dieu m'a offert, à qui je dois ma réussite et tout mon respect, mon cher père ma source de vie et d'affection qui a toujours été là, Que dieu te garde pour nous !

.

A ma très chère mère BEN EL HADJ Hadjila,

Aucune dédicace ne saurait exprimer l'amour, l'estime, le dévouement et le respect que j'ai toujours eu pour toi, et rien au monde vaut les efforts fournis jour et nuit pour mon éducation et mon bien être, tu as toujours été présente à mes côtés et me consoler quand il fallait tu n'as cessé de me soutenir, Sans ton aide, tes conseils et tes encouragements ce travail n'aurait pas vu le jour.

En cette occasion aussi mémorable, pour moi ainsi que pour toi, reçoit ce travail (fuit de tes sacrifices que tu as consentis pour mon éducation et ma formation) en signe de ma vive reconnaissance et ma profonde estime. Puisse le tout puissant te donner santé, bonheur et longue vie afin que je puisse te combler à mon tour. Sans vous deux je n'existe pas

A mes très chères soeurs Djouher / Lilia / Malika,

Mes chères soeurs, les mots ne suffisent guère pour exprimer l'attachement, l'amour et l'affection que je porte pour vous, mes anges gardiens et mes fidèles accompagnantes dans les moments les plus délicats de cette vie, je suis chanceuse de vous avoir à mes côtés. Je vous dédie ce travail avec tous mes voeux de bonheur, de santé et de réussite.

A mon cousin BEN AMI Achour et sa famille,

Qui m'est un grand frère pour moi je te remercie du plus profond de mon coeur pour ta présence et tes encouragements toi et ta femme ***Zahoua*** *pour mon parcours d'études, que dieu te garde pour tes enfants à qui je souhaite la réussite dans leurs parcours. Et je dédie mon travail à toute la famille* ***BENAMI.***

A mes très chers meilleurs amis Taoues , ***Yahia , Amal , Kahina , Sarah,*** *et* ***Rafik ainsi qu'à leurs familles au complet,***

Qui sont soeur mon bras droit que j'aime tant qui ont contribué à la réalisation de ce travail, Je vous remercie pour votre amitié chère à mon cœur, et ta présence à mes côtés dans les moments difficiles et pour vos encouragements je vous souhaite une bonne continuation dans vos vies.

Je vous aime tous!

BENAMI Kenza

Dédicaces

Du plus profond de mon cœur je dédie ce travail à ceux qui me sont chers tous mes ami(e)s sans exception,

A mon très cher père,

A l'homme qui m'est précieux, la meilleure offre que le bon Dieu m'a offert, à qui je dois ma réussite et tout mon respect, mon cher père ma source de vie et d'affection qui a toujours été là, Que dieu te garde pour nous.

A ma très chère mère,

Aucune dédicace ne saurait exprimer l'amour, l'estime, le dévouement et le respect que j'ai toujours eu pour toi, et rien au monde vaut les efforts fournis jour et nuit pour mon éducation et mon bien être, tu as toujours été présente à mes côtés et me consoler quand il fallait tu n'as cessé de me soutenir, Sans ton aide, tes conseils et tes encouragements ce travail n'aurait pas vu le jour.

En cette occasion aussi mémorable, pour moi ainsi que pour toi, reçoit ce travail fruit de tes sacrifices que tu as consentis pour mon éducation et ma formation, en signe de ma vive reconnaissance et ma profonde estime. Puisse le tout puissant te donner santé, bonheur et longue vie afin que je puisse te combler à mon tour. Sans vous deux je n'existe pas.

A celui que j'aime beaucoup et qui m'a soutenue tout au long de ce projet : mon fiancé ***YAKOUB Akli*** *et ma belle-mère Nacira, mon beau-père Hamid et ses frères: Sofiane et Ghiles, Merci pour vos encouragements et grâce à vous j'ai pu surmonter toutes les difficultés.*

A ma très chère soeur Souhila et mes deux frères Moumouh / Nadir:

Mes chèr(e)s, les mots ne suffisent guère pour exprimer l'attachement, l'amour et l'affection que je porte pour vous trois, mes anges gardiens et mes fidèles accompagnants dans les moments les plus délicats de cette vie, je suis chanceuse de vous avoir à mes côtés. Je vous dédie ce travail avec tous mes voeux de bonheur, de santé et de réussite

Dédicace à toutes mes tentes et oncles et à la famille BEN OUALI en général.

À ma meilleure amie, et binôme Kenza et tous mes chers ami(e)s

Je ne peux trouver les mots justes et sincères pour vous exprimer mon affection et mes pensées, vous êtes ceux sur qui je peux compter. En témoignage de l'amitié qui nous uni et des souvenirs de tous les moments que nous avons passés ensemble, je vous dédie ce travail et je vous souhaite une vie pleine de santé et de bonheur.

Je vous estime tous … !

BEN OUALI Taoues

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) dans les microblogs plus précisément dans Twitter.

De nos jours, les réseaux sociaux (RS) représentent le moyen le plus utilisé pour la communication, le partage de connaissances et de contenus sur le web. Avec cette dimension sociale qui enrichit les contenus des ressources, les besoins en information changent. D'où l'émergence de la RI Sociale (RIS), une thématique récente qui prend en compte les informations spécifiques aux RS ce qui implique l'occupation des microblogs d'une grande part de l'information générée sur le web.

Notre objectif est d'améliorer le score thématique dans la recherche de microblogs plus précisément dans Twitter, pour ce faire, nous avons proposé une approche qui exploite de plus des signaux sociaux de Twitter le facteur temps.

Les mots clés : recherche d'information, recherche d'information sociale , temporalité , microblogs , Twitter , signaux sociaux , commentaires , retweets .

Abstract

Our work takes place in the context of social information retrieval (RIS) and is particularly interested in the exploitation of social cues in the process of information retrieval (IR) in microblogs more specifically in Twitter.

Nowadays, social networks (RS) represent the most used means for communication, sharing of knowledge and content on the web. With this social dimension that enriches the content of resources, information needs change. Hence the emergence of Social IR (RIS), a recent theme that takes into account information specific to SRs which implies the occupation of microblogs of a large part of the information generated on the web.

Our goal is to improve the thematic score in microblogging research specifically in Twitter, to do this we have proposed an approach that additionally exploits social signals from Twitter on the time factor.

The keywords: information research, social information research, temporality, microblogs, Twitter, social signals, comments, retweets.

Sommaire

Chapitre III : Etat de l'art de la recherche d'information dans Twitter

Chapitre IV : Proposition et expérimentation de l'approche

Tables de figures

Liste des tableaux

Introduction générale

Contexte et problématique

Au début le web est composé de pages statiques reliées entre elles par des hyperliens, et s'est rapidement orienté vers un cadre collaboratif, où tous les internautes consultent, créent, partagent et diffusent de l'information.

L'évolution du web a remis la RI face à de nouveaux défis, le but est de retrouver une information pertinente dans une grande collection. Ce qui a donné naissance à la recherche d'information sociale (RIS). Celle-ci donc est un nouveau paradigme de recherche. Elle consiste à adapter les modèles et les algorithmes de la RI classique en exploitant les réactions des utilisateurs sur les ressources du Web. La motivation derrière l'exploitation des 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.

Des travaux tentent d'utiliser différents facteurs tirés des plateformes de microblogging, informations sociales, afin d'améliorer les modèles développés pour la RI classique. Dans notre cas nous avons cherché à explorer le facteur temporel en plus du contenu des tweets pour tenter d'améliorer le score thématique de la recherche. Pour cela nous avons proposé une approche qui utilise le facteur temps et le nombre de commentaires.

Organisation du mémoire

Notre travail est organisé selon la structure suivante:

Chapitre I: ***«Généralités sur la recherche d'information»*** dans ce chapitre nous passons en revue les concepts généraux de la recherche d'information (RI) . Par la suite, nous présentons le processus en «U» de la (RI) en détaillant ses étapes. Enfin , nous concluons avec les principaux modèles de la RI ainsi que les mesures d'évaluations des systèmes de recherche d'information (SRI).

Chapitre II: ***«La RI sociale et la RI dans les microblogs»*** à ce niveau , nous structurons ce chapitre en deux parties; la première traite la recherche d'information sociale (RIS) où nous avons abordé les types de cette dernière sur internet et son exploitation . La seconde partie, traite la recherche d'information dans les microblogs en particulier la plateforme Twitter.

Chapitre III: ***«L'état de l'art de la recherche d'information dans Twitter»*** nous présentons un aperçu sur les travaux de l'état de l'art consacrés à l'exploitation des informations sociale et du facteur temporel dans le processus de la RI

Chapitre IV: **«*Proposition et expérimentation de l'approche*»** ce chapitre présente notre contribution qui se base sur l'intégration et l'exploitation des signaux sociaux de Twitter et de temporalité. Où nous présentons l'approche que nous avons proposée ainsi que les détails de son implémentation et de sa mise en œuvre, ainsi que les résultats de son évaluation.

Enfin, nous terminons notre mémoire par une conclusion générale, des perspectives et des propositions.

CHAPITRE I

LA RECHERCHE D'INFORMATION (RI)

Introduction

Le domaine de la RI remonte au début des années 1950, peu après l'invention des ordinateurs [1]. Comme plusieurs autres domaines informatiques, les pionniers de l'époque étaient enthousiastes à utiliser l'ordinateur pour automatiser la recherche des informations notamment dans les bibliothèques.

Le nom de « recherche d'information » (***information retrieval***) fut donné par Calvin N. Mooers en 1948 pour la première fois quand il travaillait sur son mémoire de maîtrise [1] . La première conférence dédiée à ce thème (*International Conference on Scientific Information)* s'est tenue en 1958 à Washington. On y comptait les pionniers du domaine, notamment, Cyril Cleverdon, Brian Campbell Vickery, Peter Luhn, etc.

I. 1 Système de recherche d'information

Un système de recherche d'information est un système qui permet de retrouver les documents pertinents à une requête d'un utilisateur, à partir d'une base de documents volumineuse. [A]

Dans cette définition, il y a trois notions clés: documents, requête, pertinence.

Document: toute unité qui peut constituer une réponse à une requête d'utilisateur. Peut-être un texte, page Web, image, vidéo... etc.

Requête: est un ensemble de mots-clés, c'est l'expression du besoin en information d'un utilisateur. Elle est en général formulée en langage naturel en spécifiant une expression particulière.

Pertinence: le but de la RI est de trouver seulement les documents pertinents. La notion de pertinence est très complexe. De façon générale, dans le document pertinent, l'utilisateur doit pouvoir trouver les informations dont il a besoin. C'est sur cette notion de pertinence que le système doit juger si un document doit être donné à l'utilisateur comme réponse, donc c'est le degré d'utilité du document pour l'utilisateur.

Vu que toutes les évaluations de la RI ne tournent qu'autour de la notion de pertinence donc elle est l'élément pivot de celle-ci. La pertinence est perçue sur les deux niveaux ;

- **le niveau utilisateur :** à ce niveau, l'utilisateur a un besoin d'information, et il espère obtenir les documents pertinents pour satisfaire ce besoin. La relation entre le besoin d'information et les documents attendus est la relation de pertinence.

- **le niveau système**: est défini à travers les modèles de la recherche d'information et est traduite par un score évaluant l'adéquation du contenu des documents vis-à-vis de celui de la requête.

I.2 Processus en «U» de la Recherche d'Information

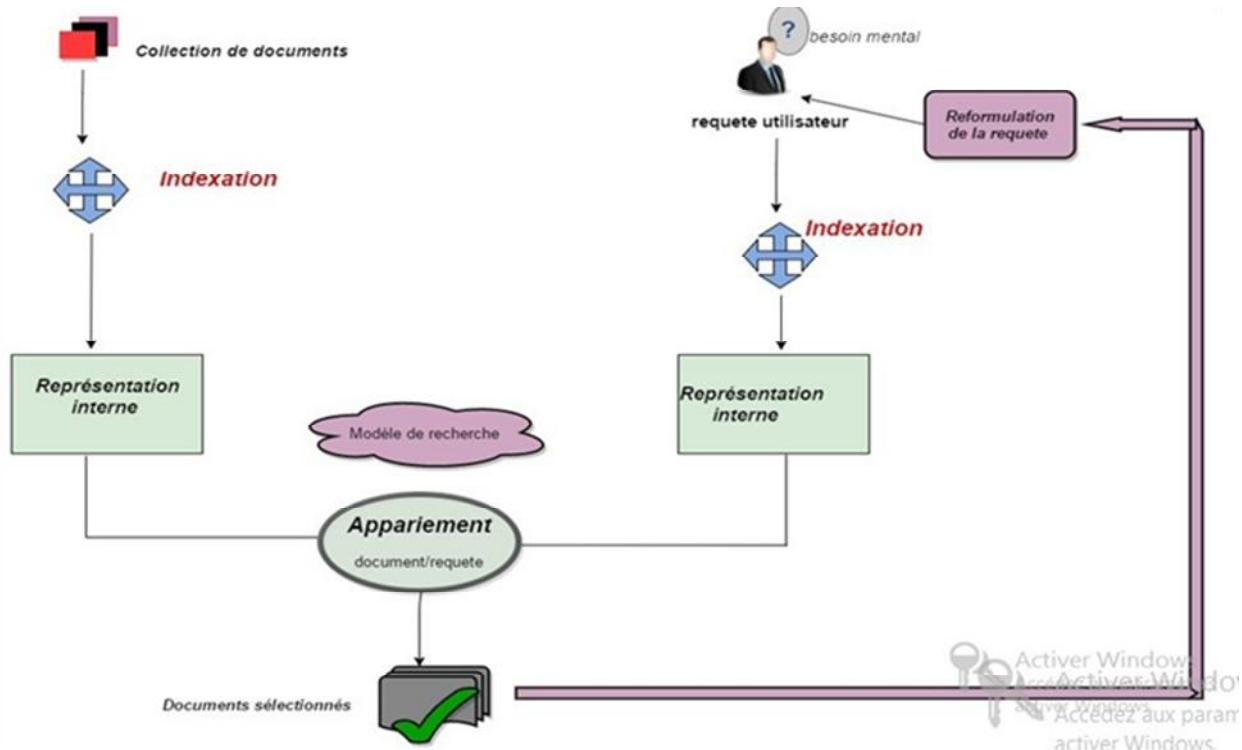


Figure (1) : -Processus en «U» de la Recherche d'Information -

Afin de réaliser une certaine correspondance entre des informations contenues dans la collection de documents textuels ainsi que le besoin en information d'un utilisateur autrement dit la requête un SRI (système de recherche d'information) met en œuvre le processus ci-dessus qui intègre les fonctions qui suivent :

I.2.1 L'indexation

Le processus d'indexation consiste à extraire des documents les termes (ou concepts) jugés les plus significatifs et pertinents afin d'en construire une représentation médiatrice entre les documents et les utilisateurs. Il s'agit de décrire leurs contenus et de les représenter par des index. Elle peut être:

- **Manuelle**: le choix des termes d'indexation et la définition des mots clés représentatifs du contenu du document dépendent de l'indexeur et de ses connaissances du domaine.
- **Automatique :** c'est un processus complètement automatisé qui se charge d'extraire les termes caractéristiques du document.
- **Semi-automatique** *:* est une combinaison des deux approches d'indexation manuelle et automatique.

Une indexation automatique est d'abord lancée, qui extrait un des termes descripteurs du document, suivie par l'indexation manuelle où l'indexeur humain fait le choix final des termes à partir du vocabulaire fourni.

Durant notre travail nous allons spécialement nous intéresser sur l'indexation ***automatique***.

Celle-ci consiste à extraire automatiquement des termes du document, à éliminer des mots vides, à faire la normalisation, la pondération et pour enfin créer l'index en question

Étapes suivies pour une indexation automatique

1. **L'extraction automatique des termes**

Si on dispose d'une collection de documents en divers formats (html, texte brut, pdf, word, open office …etc) , diverses langues et encodages, on en prend quelques-uns contenants des textes :

En faisant d'abord, de la conversion des majuscules en minuscules, puis l'élimination des accents, et enfin on procède à la tokénisation qui est un processus de conversion d'un texte en une séquence d'unités lexicales élémentaires.

2. **L'élimination des mots vides**

Les mots vides sont tous les mots qui n'apportent pas de sens au texte et qui ne traitent pas le sujet d'un document tels que :***les déterminants** : le, la, les.. ***les pronoms** : je, tu … ***les prépositions** : sur, contre, à, pour, de … ***Les adverbes** : comme, bientôt, ensuite …***Certains verbes fonctionnels** : être, vouloir, avoir, …

L'objectif est d'identifier ces mots vides et de les exclure de l'index en utilisant une liste prédéfinie de ces derniers appelée (**stop-list** ou un anti-dictionnaire) ; cela permet d'économiser beaucoup d'espace mémoire d'exécution ainsi de réduire l'index. Mais attention parfois ils sont porteurs de sens.

3. **La normalisation**

On peut trouver de diverses formes d'un mot dans un texte toutes relatives à un même sens on n'a qu'à représenter ces dernières par une unique racine grammaticale qui sera capable de représenter le concept véhiculé et de ramener les mots de la même famille à leur forme normale et ça par flexion (--Verbale – Nominale – forme canonique), dérivation (ajout de suffixes ou préfixes)…

4. La pondération

Elle mesure l'importance d'un mot dans un document (d'une collection) donné, afin de caractériser son influence sur la représentation de document, en lui associant une valeur numérique qui représente cette importance-là. Cette dernière est calculée en utilisant des approches basées sur des aspects statistiques. Nous en citant :

***Approches basées sur la fréquence locale**

Est la fréquence d'occurrence des mots dans un document et est notée ***Tf***. Le but est de trouver les mots qui représentent le mieux le contenu d'un document .C'est celui qui apparaît souvent dans le texte qui représente un concept important.

***Approches basées sur la fréquence globale**

Notée ***Idf***, on peut dire que c'est la pondération globale ; elle mesure la fréquence d'un terme dans une collection de documents. L'approche dit que si un terme est présent dans plusieurs documents a moins d'importance qu'un terme moins fréquent dans la collection vu qu'il ne caractérise aucun document en particulier.

La formule qui décrit cette mesure est la suivante :

$$Idf_i = log(\frac{N}{n+1}) \qquad \textbf{(1.1)}$$

Où:

- **t**: est le terme
- **N** : est le nombre de documents dans la collection
- **n** :est le nombre de documents où t apparaît

***Approches combinées**

La fonction de pondération combinant la pondération locale et globale est référencée sous le nom de la mesure ***tf*idf ,*** qui donne une bonne approximation de l'importance du terme dans les collections de documents. Exprimée par cette formule :

$$TfIdf = Tf_{t,d} \times Idf_t \qquad (1.2)$$

Enfin on procède à la création de l'index. Ce dernier est représenté par des structures de données tels qu'une matrice d'incidence (termes-documents) [contenant des 1 si le terme figure dans le document et des 0 sinon], ainsi par un fichier inverse contenant le vocabulaire et ses occurrences.

I.2.2 L'appariement requête / document

Le SRI prédit les documents que l'utilisateur trouvera pertinent

* En faisant la correspondance de la requête et de l'index,
* Puis calcule un score de pertinence qui reflète le degré de similarité entre la requête et le document. Ce score est calculé à partir d'une valeur ***RSV(q, d)***

(Retrieval Status Value),

Où : ***q*** est une requête

D est un document.

Cette mesure tient en compte la pondération du terme.

I.2.3 La reformulation de la requête

Des fois, les résultats que fournit le système de recherche d'information ne satisfont pas le besoin d'information de l'utilisateur, dû à la mal expression de la requête par ce dernier.

Pour correspondre la pertinence utilisateur et la pertinence système , le SRI passe @la reformulation ou bien la modification de la requête initiale par l'ajout de termes significatifs ou de ré-estimation de leurs poids.

Cette étape peut être effectuée soit **manuellement** et cela par la soumission d'une nouvelle requête par l'utilisateur ; ou **dynamiquement** par injection de pertinence basée sur les jugements de l'utilisateur sur des documents restitués par le système comme réponse à la requête initiale. Cette méthode consiste en la sélection des termes importants appartenant aux documents jugés pertinents par l'utilisateur, et leur exploitation dans la formulation de la nouvelle requête.

I.3 Modèles de recherche d'information

Un Modèle de recherche d'information est capable de créer une représentation interne pour un document ou pour une requête basée sur les termes pondérés issus de l'indexation ou encore, de définir une méthode de comparaison entre une représentation de document et une représentation de requête pour déterminer leur degré de similarité.

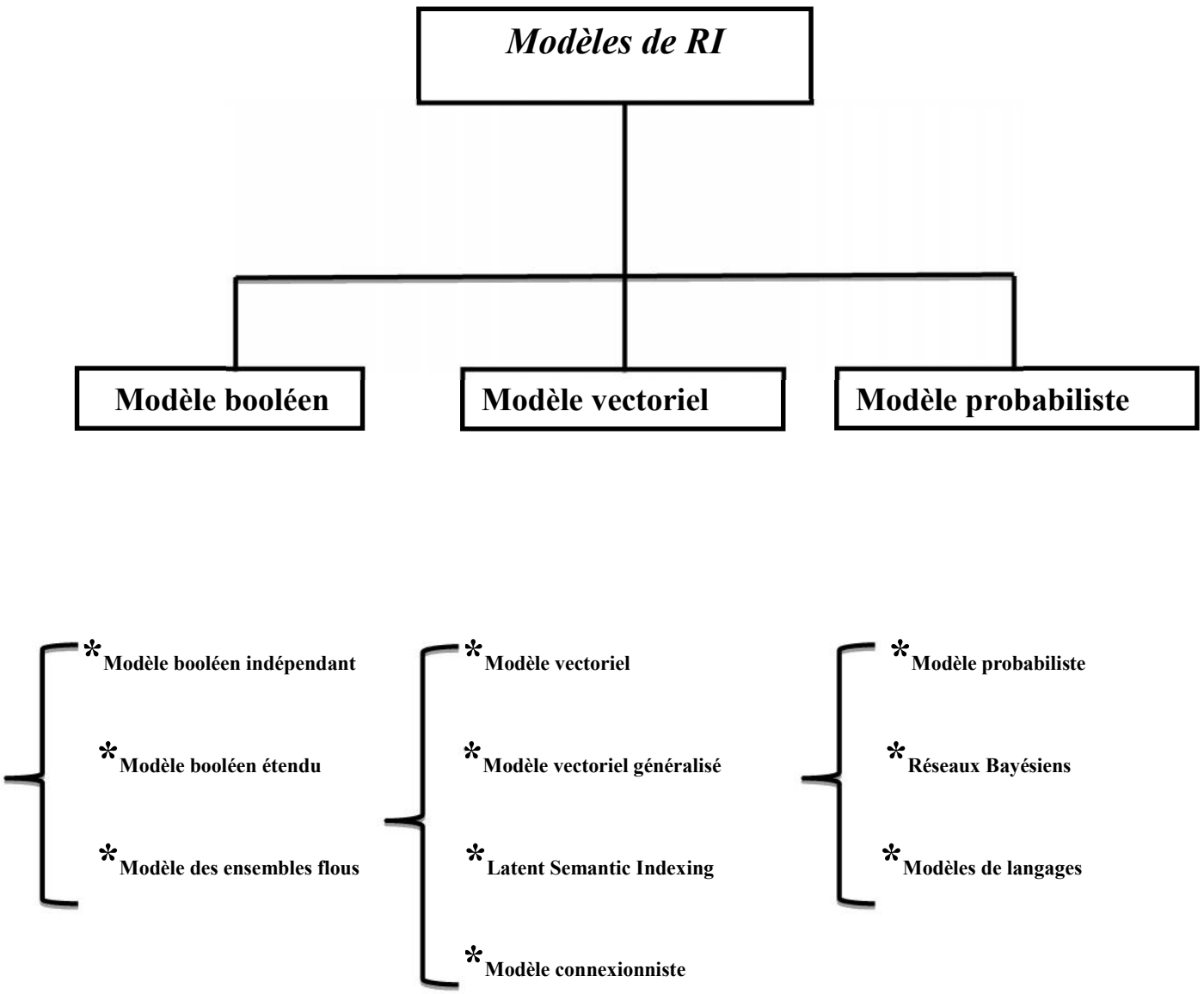


Figure (2) : - le schéma des modèles de la RI classique. - [2]

I.3.1 Le modèle booléen

C'est le modèle le plus simple de la RI et est basé sur la théorie des ensembles et l'algèbre booléenne. [3] .Le document est représenté par une conjonction de termes non pondérés qui constitue l'index de ce dernier, et la requête est un ensemble de mots clés reliés par des opérateurs booléens (AND, OR et NOT).

L'appariement requête-document est strict et se base sur les règles suivantes :

RSV (d,t_i)=1 si ti□ d ; 0 sinon **(1.3)**

RSV (d , q_i AND q_j) =1 si ***RSV (d, q_i)*** =1 et ***RSV(d,q_j)*** =1 ; 0 sinon.

RSV(d , q_i OR q_j) =1 si ***RSV (d, qi)*** =1ou ***RSV(d,q_j)***=1 ; 0 sinon.

RSV (d, NOT q_i) =1 si ***RSV (d, q_i)*** = 0 ; 0 sinon.

Où : ***RSV*** est la fonction de similarité entre une requête **q** ou un terme **t** et un document **d**

La fonction de pertinence ***RSV (d , q)*** 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 (d , q)*** =1 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 :

* L'absence de classement des documents : en effet, le système retourne un ensemble de documents non ordonné à une requête. Du coup, si la liste retournée est longue l'utilisateur sera obligé de parcourir toute cette longue liste non ordonnée afin d'identifier le document pertinent à sa requête.

* Et l'expression logique devient compliquée quand la requête est longue.

I.3.2 Le modèle vectoriel

Le modèle vectoriel [4] est un modèle basé sur l'algèbre. Le document est représenté par un vecteur de poids w_{ij} de dimension n, dans l'espace vectoriel composé de tous les termes d'indexation de même dimension. La requête est représentée par un vecteur de poids w_{Qj} de même dimension que celui du document. La pertinence du document pour la requête est mesurée comme le degré de corrélation des vecteurs correspondants avec les mesures suivantes:

1*Le produit scalaire

$$Sim\ (d, Q) = \sum_{j=1}^{n} w_{Q_j} \square w_{ij} \quad [5] \qquad (1.4)$$

2*La mesure du cosinus

$$Sim\ (d\ ,\ Q) = \frac{\sum_{j=1}^{n}(w_{Q_j} \square w_{ij})}{(\sum_{j=1}^{n}(w_{Q_j})^{1/2}) \square (\sum_{j=1}^{n}(w_{ij})^{1/2})} \quad [5] \qquad (1.5)$$

3* La mesure de Jaccard

$$Sim\ (d,Q) = \frac{\sum_{j=1}^{n}(w_{Q_j} \square w_{ij})}{(\sum_{j=1}^{n} w_{Q_j}{}^{2}) + (\sum_{j=1}^{n} w_{ij}^{2}) * \sum_{j=1}^{n}(w_{Q_j} \square w_{ij})} \qquad (1.6)$$

Et bien d'autres mesures.

Contrairement au modèle précédent le modèle vectoriel, il trie les résultats selon leurs niveaux de pertinence vis-à-vis de la requête et ça grâce aux mesures de corrélation vectorielles.

Mais le problème qu'on a rencontré dans ce modèle est l'indépendance entre les termes d'un même document ou requête alors que les termes dans le document sont sémantiquement liés.

I.3.3 Le modèle probabiliste

Le modèle probabiliste [6] consiste à présenter les résultats du SRI dans un ordre basé sur la probabilité de pertinence d'un document par rapport à une requête.

Le modèle tente d'estimer la probabilité qu'un document **d** appartienne à la classe des documents pertinents /non pertinents

Le SRI retourne un document **d** si : $P(R/d) > P(\neg R/d)$

Où:

- **d** est le document,
- ***R*** est l'ensemble des documents pertinents,
- ***¬ R*** est l'ensemble des documents non pertinents.
- ***P(R/d)*** est la probabilité que le document **d** appartienne aux documents pertinents pour la requête Q,
- ***P(¬ R /d)*** est la probabilité que le document **d** appartienne aux documents non pertinent pour la requête Q,

La formule RSV (d,Q) est donnée après simplification et avec l'utilisation de la formule de Bayes ainsi :

$$RSV(d, Q) = \frac{P(d/R)}{P(d/\neg R)} \qquad (1.7)$$

P(d/R)(respectivement ***P(d/ ¬ R)***) est la probabilité que le document d appartienne à

R l'ensemble des documents pertinents (respectivement à ¬ ***R*** l'ensemble des documents non pertinents) afin d'estimer ces probabilités on a utilisé plusieurs méthodes parmi ces dernières la plus connue est celle du modèle de ***BIR*** (Binary Independance Retrieval) où:

Le document d est une variable représentée par un ensemble d'événements d (t1=**x1**, t2=**x2** , t3=**x3**,…, tn=**xn**)

indépendants qui dénotent la présence si **xi=1** ou bien l'absence si **xi=0** d'un terme **t** dans le document **d**

Alors les probabilités P(d/R) et P (d/ ¬ *R*) sont calculées par :

$$P(d/R)=\prod_{i=1}^{i=n} P(ti = x/R) \qquad \textbf{(1.8)}$$

$$P(d/R)=\prod_{i=1}^{i=n} P(t_i = x/\neg R) \qquad \textbf{(1.9)}$$

La fonction RSV(d,Q) sera écrite après la transformation ainsi :

$$RSV(d,Q) = \sum_{t_i \square T} x \square log\left(\frac{1-P(t_i D\backslash R)}{(1-P(t_i D\backslash \neg R)}\right) \qquad \textbf{(1.10)}$$

n posant **:** $P(ti\square D\backslash R)=\boldsymbol{pi}$ et $P(ti\square D\backslash\neg R)=\boldsymbol{qi}$

On aura: $\sum_{t_i \in T} x \log\left(\frac{1-p_i}{1-q_i}\right)$ **(1.11)**

L'estimation des probabilités se fait si l'ensemble R(respectivement¬ *R*) est connu ce qui fait l'inconvénient de ce modèle , pour y remédier un autre modèle est venu, nommé 2-poisson proposé par S.Robertson qui a abouti à une formule appelée (BM25) qui inclut la fréquence locale d'un terme dans un document ainsi la longueur de ce dernier.

$$\boldsymbol{RSV}(\boldsymbol{d},\boldsymbol{Q}) = \sum_{t_j \in Q} \boldsymbol{tf} \log\left(\frac{\boldsymbol{N}-\boldsymbol{df_j}+0.5}{\boldsymbol{df_j}+0.5}\right) * \frac{\boldsymbol{k_1}+1}{\boldsymbol{k_1}\left(1-\boldsymbol{b}+\frac{\boldsymbol{b}*\boldsymbol{l_j}}{\boldsymbol{avgdl}}\right)+\boldsymbol{tf_{ij}}} \quad \textbf{(1.12)}$$

d : est le document

Q : est la requête

tj : est le termej

tfj: est la fréquence de tj dans la requête Q

N : le nombre de documents dans le corpus

Où:

df_j: est le nombre de documents contenant le terme tj

$k_1$et ***b*** : sont deux constantes qui dépendent de la collection et du type des requêtes

l_i: est la longueur du document d

avgdl: est la moyenne des longueurs des documents dans le corpus

tf_{ij}: est la fréquence de $t_j$dans le document d_i

I.4 L'évaluation des SRI

L'évaluation du SRI [7] constitue une étape primordiale durant la mise en œuvre d'un modèle de recherche d'information. Elle permet de paramétrer le modèle ainsi d'estimer l'impact de toutes ses caractéristiques et aussi de fournir des éléments de comparaison entre modèles. L'évaluation du SRI se mesure en s'appuyant sur des techniques basées sur l'estimation de la qualité des résultats d'un SRI. La qualité de ce dernier est mesurée par la comparaison des résultats avec les réponses qui satisfont le besoin de l'utilisateur. Un système se dit performant si ses réponses correspondent à celle que l'utilisateur espère recevoir, ainsi que ceux qui sont rapides et moins gourmands de l'espace mémoire.

I.4.1 les mesures d'évaluation

Il existe deux groupes de mesures d'évaluation [B] qui permettent d'évaluer un SRI; les mesures non-ordonnées et les mesures ordonnées.

I.4.1.1 Mesures non-ordonnées

Ce groupe de mesures prend en compte uniquement le nombre de documents pertinents retournés lors de la recherche, elles ne considèrent pas l'ordre d'apparition des résultats. Les deux mesures principales sont la précision et le rappel.

La précision: est la capacité d'un système à nous sélectionner que les documents pertinents [B] . Si la précision est égale à 1 donc le SRI ne retourne que les documents pertinents

$$Précision = \frac{documents\ pertinents\ restitués}{documents\ restitués} \qquad (1.13)$$

Le rappel: est la capacité d'un système à sélectionner tous les documents pertinents de la collection..[B]

$$Rappel = \frac{documents\ pertinents restitués}{documents\ pertinents} \qquad (1.14)$$

➔ Plus le rappel est proche de 1, meilleure est la réponse du SRI.

Les mesures de rappel et de précision utilisées seules ne sont pas de bons indicateurs de la performance d'un SRI. Plusieurs approches proposées dont : Agrégation du rappel et de la précision dans une seule mesure qui est le F-Score (ou F-mesure)

F-Score(ou F-mesure) :c'est une mesure qui combine la précision et le rappel, nommée F-XCscore introduite dans (Rijsbergen, 1979) et se calcule ainsi:

$$F\_mesure = \frac{2 \times Précision \times Rappel}{Précision + Rappel} \quad \textbf{(1.15)}$$

I.4.1.2 Mesures ordonnées

Ce second groupe de mesures, prend en compte l'ordre des résultats contrairement au précédent. Les mesures sont affectées par l'ordre des documents retournés, parmi ces mesures : la précision@R, la précision moyenne et la R- précision.

- **Précision @R** : c'est la précision à différents niveaux de coupe [B]. Cette dernière mesure la proportion des documents pertinents retrouvés parmi les R premiers documents retournés par lesystème.
- **R-précision** : cette précision mesure la proportion des documents pertinents retrouvés après que R-documents ont été retrouvés, où R est le nombre de documents pertinents pour la requête considérée.
- **La précision moyenne :** (AveragepecisionAVGp) : c'est la moyenne des valeurs de précision après chaque document pertinent, elle se calcule comme suit:

$$AVGp = \frac{1}{R} \sum_{i=1}^{N} p(i) \times R(i) \quad \textbf{(1.16)}$$

Où **:**

- **R(i) =1** si : le i ème document restitué est pertinent, **R(i) =0**sinon.
- **p(i) :** la précision à i documents restitués.
- **R** : le nombre de documents pertinents pour la requête **q**
- **N :** le nombre de documents restitués par le système.

La MAP (Mean Average Precision) : c'est la moyenne des précisions moyennes obtenues sur les l'ensemble des requêtes à chaque fois qu'un document pertinent est retrouvé ($|Q|$).

$$MAP = \frac{\sum_{q \in Q} AVGp}{|Q|} \quad \textbf{(1.17)}$$

I.4.2 Collection de référence et de tests(TREC)

Une collection de tests ou bien de référence est un ensemble de requêtes-tests muni de réponses idéales associées à chaque requête utilisée dans l'évaluation du SRI. Pour cela, nous soumettons cette collection de requêtes-tests au SRI ainsi de comparer les réponses de celui-là à celles espérées dites requêtes types, comme ça nous allons obtenir une mesure de qualité sur les performances de notre SRI.

La collection la plus connue en RI est la collection **TREC** apparue dans les années 90 initiée par NIST (National Institute of Standards and Technology) et par DARPA(Defence Advanced Research Project Agency). Ce programme offre des moyens d'évaluation des SRI comme des collections de tests, protocoles d'évaluation … . Son but est de proposer un standard pour comparer plusieurs SRI ou des modèles implémentés par ces derniers et de mesurer leurs efficacités.

Conclusion

Dans ce chapitre nous avons présenté la recherche d'information classique et l'avons illustrée par une définition puis, nous y avons décrit l'architecture de son processus. Nous avons pu aborder les étapes de l'indexation. Par la suite, nous avons entamé les divers modèles de la RI, et avons su évaluer la performance d'un SRI.

La recherche classique ne s'intéresse qu'aux données textuelles. Cependant, avec l'émergence du web, plus précisément l'apparition des réseaux sociaux, l'enjeu 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, commentaires …), la temporalité, etc. Ce qui est détaillé dans le prochain chapitre qui porte sur la recherche d'information sociale (RIS).

Chapitre II

La RI sociale et la RI dans les microblogs

Introduction

La recherche d'information 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).

Nous consacrons ce chapitre à la recherche d'information sociale (RIS) , où nous allons aborder en détails juste après la RIS , spécialement dans la plateforme Twitter et nous familiariser avec cette dernière.

II.1 La recherche d'information sociale

Comme étant l'une des variantes de la RI , la RIS est une thématique récente qui a pour objectif de prendre en compte les informations spécifiques aux RS appelés les signaux sociaux qui représentent des informations communicatives et informatives ,fournissant directement (tel que: le contenu généré par l'utilisateur , blogs, forums…etc) ou indirectement (comme les liens sociaux, les profils ainsi que leurs traces de navigations.) 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).

II.1.1 Les types d'informations sociales sur internet

Comme nous l'avons cité auparavant, les utilisateurs génèrent eux même des informations, ou bien elles sont générées indirectement et cela par l'extraction de leurs comportements par internet (les traces de leurs visites des sites).

II.1.1 .1 Le contenu généré par l'utilisateur (User Generating Content)

Désigne tout type de contenus 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 tags, commentaires, ou avis.

Nous définissons ci-après ces moyens de production collaboratifs :

***Blog** : est un site web sur lequel un internaute publie des articles dits billets ou bien posts. Il s'agit d'un espace individuel d'expression, créé pour donner la parole à tous les internautes, permettre à tous les visiteurs de réagir sur le sujet évoqué, et cela en commentant les articles publiés, ce qui crée une relation privilégiée entre auteurs et lecteurs.

** **Wiki :** est un site web dynamique où l'information est construite avec la participation de plusieurs personnes, tout utilisateur peut créer, modifier et supprimer des contenus de manière collaborative, chaque modification est sauvegardée et les versions historiques restent toujours accessibles.

** **Forum :** est un espace d'échange d'informations où les internautes posent ou répondent à une question donnée. Les différentes contributions forment un fil de discussion. Les forums sont classés par thèmes bien précis. Les messages publiés dans les forums par les internautes sont archivés. Ce qui leur permet d'y participer d'une manière asynchrone.

****Microblog :** est un type de blog dans lequel les utilisateurs peuvent publier de petits morceaux de contenu numérique comme les images , les vidéos , ou l'audio sur internet ces publications appelées micro-messages, sont immédiatement accessibles à une petite communauté ou au grand publique le point de différence entre eux est la longueur du contenu qui est plus petite.

** **Les réseaux sociaux numériques :** un réseau social numérique est un site internet qui permet aux internautes de créer une page personnelle pour partager et échanger des informations, des médias avec leurs communautés d'amis ainsi leur réseaux de connaissance qui les réunissent via des échanges personnalisés, chacun peut lire les messages de tel ou tel autre utilisateur.

II.1.1 .2 Le contenu généré par la pratique

Comporte des informations communicatives qui fournissent indirectement des renseignements sur les interactions, les émotions, les relations et les comportements sociaux. On en cite [8] :

****Les traces des utilisateurs** : elles peuvent déterminer les préférences des utilisateurs et leurs thématiques de recherche et cela à travers les diverses pages web visitées par les internautes, les clicks, les durées de visites …

****Les données personnelles** : elles se composent des informations que l'utilisateur fournit lors de son inscription sur les réseaux sociaux.

****Les liens sociaux** : la plupart des plateformes sociales définissent des règles de liaison entre leurs différents utilisateurs. Ces dernières diffèrent d'une plateforme à une autre. On a celles qui n'admettent pas de restriction dans les liens sociaux (comme Twitter) à moins que le compte soit privé, contrairement à d'autres comme Facebook, où les deux utilisateurs doivent être d'accord pour partager leurs informations.

II.1.2 Exploitation des informations sociales

La RI sociale consiste à adapter les modèles de la RI classique en exploitant les informations sociales, par exemple les connaissances des utilisateurs experts ou bien les expériences de recherche des autres utilisateurs. Cela en considérant les annotations sociales (Peters et *al.,* 2011), l'analyse des réseaux sociaux (Kazai et Milic-Frayling, 2008), les jugements de pertinence subjectifs (Xu et al., 2007) et la recherche collaborative (Karamuftuoglu, 1998) dans le processus de la RI.

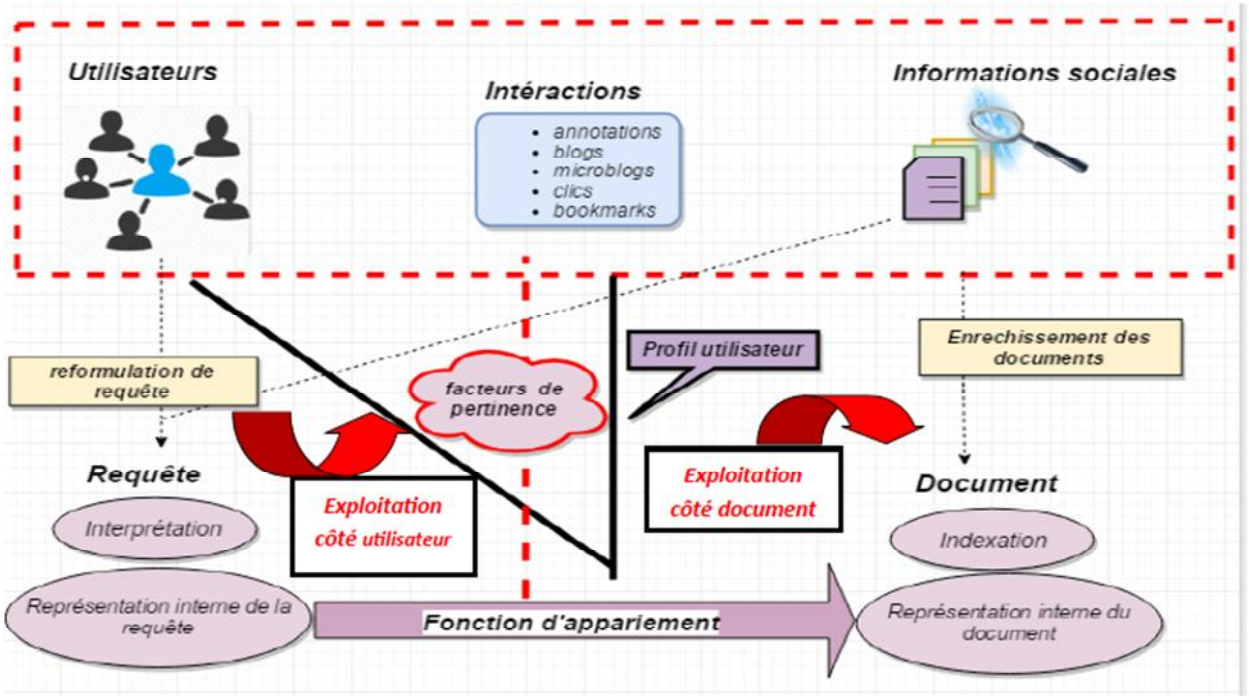


Figure 3 - Exploitation de l'information sociale dans la RI classique [8] –

Comme le montre la figure **3**, les informations sociales peuvent être exploitées au sein même du modèle de RI, et même comme une source parmi d'autres dans le web. Donc elles sont employées à plusieurs niveaux. Nous allons nous baser sur les approches qui les exploitent en fonction du niveau de leur utilisation soit du côté utilisateur ou bien du côté document.

II.1.2.1 Côté utilisateur

L'idée est d'améliorer l'efficacité de notre SRI en exploitant le contexte de l'utilisateur en tenant compte de ses informations tel que son profil dans le processus de recherche ou bien en améliorant la représentation de son besoin d'information (autrement dit la reformulation de sa requête), dans le but de retrouver des résultats plus spécifiques et plus raffinés et aussi en lui créant un profil pour une recherche personnalisée.

- **Exploitation de l'information sociale pour la reformulation de requêtes**

La reformulation de requêtes est vue comme un traitement pour élargir le champ de recherche pour celle-ci autrement dit , une requête étendue va contenir plus de termes reliés permettant d'une part de désambiguïser les mots initiaux et connaître exactement leurs sens, et d'autre part d'augmenter les chances de restituer le maximum de documents pertinents.

- **Exploitation de l'information sociale pour la création de profil et la recherche personnalisée**

Les informations sociales sont utilisées pour créer les profils des utilisateurs. Ces derniers sont en outre utilisés pour définir un contexte de restitution permettant de sélectionner des résultats personnalisés. Les éléments souvent utilisés pour créer le profil d'un utilisateur sont ses relations sociales, ses annotations et ses activités dans les plateformes sociales. Les profils à base d'informations sociales sont utilisés pour faciliter la personnalisation des recherches. Cai et Li (2010) ont proposé une approche qui permet de créer des profils d'utilisateurs basés sur les tags, ainsi que la création de profils des ressources à rechercher.

II.1.2.2 Côté documents

L'idée de l'utilisation des informations sociales comme les tags, les commentaires…etc du côté des documents est de ramener des informations supplémentaires pour enrichir la représentation des contenus recherchés ou bien pour les utiliser comme des facteurs de pertinence.

II.2 La recherche d'information dans les microblogs

Les microblogs sont de plus en plus répandus sur internet, un microblog est un blog allégé et c'est aussi une manière de publier des contenus textuels en format court (entre 140 à 200 caractères) et sans titre dans des plateformes sociales spécifiques. Ces dernières se caractérisent par l'intensité des interactions sociales entre les individus mais aussi par le flux d'information qui y circule. Le but est de partager des informations simplement et rapidement, exprimer son opinion sur n'importe quel sujet, commenter les publications des autres utilisateurs (hyperliens, des photos), il est possible de suivre de divers flux d'informations de personnes, de groupes, ainsi que des canaux de discussions.

Vu que les informations ne sont pas instantanément indexées, elles ne sont pas disponibles pour les moteurs de recherche classiques. En effet, la recherche dans les microblogs est différente de la recherche documentaire sur le web, à cause de la diversité de la structure et du format des données. De plus les motivations de recherche sont spécifiques aux microblogs, les requêtes sont souvent motivées par l'activité sociale de la personne concernée ainsi que les tendances et les événements courants.la pertinence d'une information dépend du contexte social de celle-ci, ainsi qu'une information a pratiquement la même importance que

la personne qui la publie. Il existe plusieurs plateformes de microblogging. Les cinq plateformes les plus utilisées sont : Twitter, FriendFeed , Tumblr, Posterous et Identi.ca.

➔ Nous nous focalisons sur Twitter ; un service qui permet de partager de l'information, en moins de 140 caractères, mais aussi de suivre les autres personnes, utilisé également comme source d'information.

II.2.1 La Plateforme de microblogging de Twitter

Il fut créé en 2006 aux Etats-Unis, Twitter [C] est un réseau social de microbloging qui permet aux individus d'envoyer gratuitement de brefs messages appelés tweets, par internet, par messagerie instantanée ou par SMS (messages sont limités à 280 caractères). Ces informations concernent différents sujets où des internautes parlent de leurs quotidiens, des événements, des tendances . . . etc. Twitter a connu une croissance exponentielle durant ces dernières années. Nous présentons ci-dessous les principales spécificités de cette plate-forme, ainsi que l'information qui y est produite.

II.2.2 Concepts et fonctionnement de Twitter

Afin de créer un profil Twitter on doit passer par les étapes suivantes :

1)→ Saisir l'URL de twitter : w ww.twitter.com : une page d'accueil va s'afficher , on aura le choix de se connecter si on possède déjà un compte twitter sinon on procède à la création,

2)→ Le remplissage des champs du formulaire

Figure (4). -Remplissage des champs Nom/ Prénom/ email –

3)→ La vérification des coordonnées saisies par l'utilisateur

Figure (5) . -Vérification des coordonnées -

5)→ La saisie du mot de passe :

Figure (6). - Saisir le mot de passe –

Et on peut aussi : se décrire à travers une biographie, choisir une photo pour notre profil ainsi les sujets qui nous intéressent tels que le sport, les actualités (météorologie, histoire, politique, santé . . . etc), style de vie (parentalité, bricolage et maison, voyage, fitness et bien-être.. ..etc), art et culture, jeux vidéo et bien d'autres activités.

Une fois que le profil est créé, nous pouvons commencer à tweeter, et à chaque fois qu'on le fait le message apparaît sur notre profil. Et tous nos abonnés le reçoivent dans leur time line qui se charge de les empiler dans un ordre chronologique inverse.

Figure (7). -Petit aperçu d'un profil Twitter –

* Si A est abonné à C, alors A est appelé abonné (*follower*) de C (*followee*) et reçoit automatiquement toutes les publications de C dans son **time line**.

*Les relations d'abonnement peuvent être dans un et/ou les deux sens, si C s'abonne à son tour à A. On parle dans ce cas d'une relation d'amitié.

* Si un microbloggeur diffuse un message, tous ses abonnés le reçoivent mais il peut également envoyer un message « tweet » direct et privé à l'un de ses amis. S'il le rediffuse, le message est dit «**retweet»,** ce dernier va contenir la mention **RT**. Lors de cette rediffusion, le microbloggeur peut y ajouter de l'information complémentaire.

* Un utilisateur peut en mentionner un autre dans un message (***@mention***).

*Les entreprises ou encore les sites d'information sont aujourd'hui très présents sur les plateformes de microblogging pas seulement les individus.

II.2.3 Caractéristiques de la plateforme Twitter

- ***Twitter est un système temps-réel***: un utilisateur, en accédant à sa page, reçoit en temps-réel les microblogs de ses abonnés qui défilent sur sa page par ordre chronologique inverse, (des plus récents aux plus anciens). Si à un moment donné un nouveau microblog pertinent est publié, l'utilisateur reçoit une notification pour l'afficher.
- ***Twitter possède une hétérogénéité du contenu***: en plus du contenu textuel, un tweet peut inclure de divers signes dans son tweet tels que :

* ***@ nom d'utilisateur*** : permet d'indiquer qu'on mentionne une personne particulière,

* ***#mot*** : est un hashtag. Ce dernier indique un mot important et permet de catégoriser les microblogs selon un contexte précis.

* Les microblogs contiennent aussi des ***URL*** qui renvoient vers des pages web. Ainsi que de diverses métadonnées comme :

- ***Des données de géo-localisation*** : les terminaux équipés de GPS fournissent des informations qui permettent de localiser l'endroit où le microblog a été publié.
- ***Des données d'horodatage*** : chaque microblog est caractérisé par sa date de publication. Une information utilisée pour mesurer sa fraîcheur.
- ***Les données d'auteur*** : Twitter stock le compte de celui qui a publié chaque microblog, ce qui permet de trouver les microblogs d'un auteur en particulier.
- ***Rediffusion*** : ***RT*** (retweet) indique que le message est rediffusé. Ce mécanisme permet aux utilisateurs de repartager des microblogs qu'ils trouvent intéressants parmi ceux déjà publiés.
- ***La mise en favoris*** : pour montrer son intérêt pour un tweet donné.

II.2.4 La recherche dans les microblogs de Twitter

Lors de sa recherche, un utilisateur peut mélanger entre les comptes utilisateurs, les hashtags et même des URLs et lui sera suggérées des sorties de divers types de données de recherche. Ces dernières diffèrent; si l'utilisateur sélectionne un compte utilisateur parmi la liste des suggestions, il lui sera affiché le profil de ce compte utilisateur. Sinon ça lui affichera des microblogs contenant les termes, le hashtag ou l'URL recherchée.

Une étude [8] est faite sur les motivations d'utilisation du moteur de recherche de Twitter, après une observation sur une population active sur cette plateforme, on a pu identifier ses pratiques de recherche basées sur les informations récentes comme les actualités, la tendance, les événements récents . . . etc.

Et aussi les informations sociales : les recherches portées sur les informations des utilisateurs en particulier. Enfin celles basées sur les informations sur des sujets spécifiques en relation avec les recherches effectuées sur les moteurs de recherche du Web.

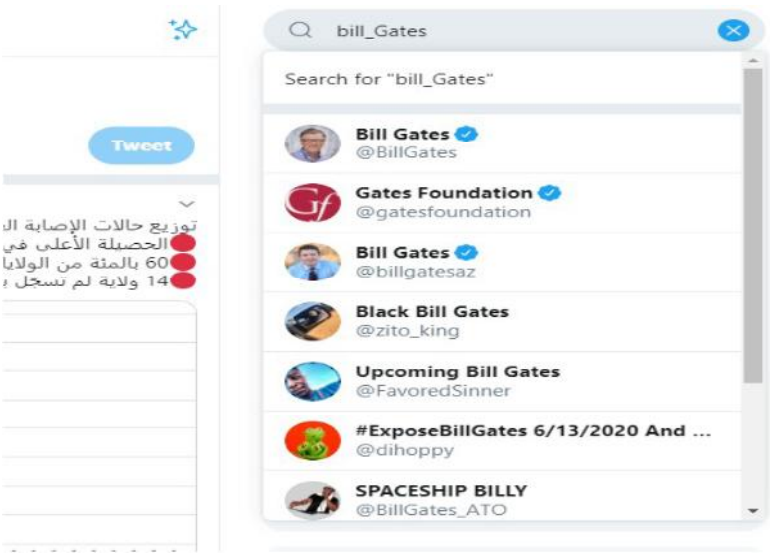


Figure (8) : -Aperçu des suggestions d'une recherche sur Twitter –

II.2.4.1 Les types d'information recherchée dans Twitter

A ce niveau nous classons les approches de recherche dans les microblogs selon le type d'information recherchée.

- ***Recherche temps réel***

Dans la RI dans les microblogs temps-réel, la date de publication d'un document est considérée comme un facteur de pertinence très important ça nous permet de trouver en temps réel les informations fiables sur des événements produits récemment, pour cela (Ounis et al.2011 [15]) a utilisé l'une des méthodes qui classe tous les résultats pertinents ainsi ceux qui sont temporellement proche à la date de soumission de la requête.

- ***Détection d'opinion***

De nombreuses recherches sont lancées dans les moteurs de recherche des plateformes de microblogging à propos des opinions claires des utilisateurs sur des produits ou bien sur un sujet en particulier (sujet de la requête).

Les microblogs expriment des opinions. Jansen a étudié l'opinion des utilisateurs sur un produit d'une certaine marque, l'étude a montré que la plupart des tweets ont un taux positif. En outre, les microblogs permettent d'obtenir des opinions et des réactions immédiates sur des produits, des marques.

- ***Détection de tendances***

La tendance est un événement de grande envergure qui intéresse le grand publique comme les élections, les événements sportifs, la mode….etc .

La détection de tendances se base sur la recherche d'expressions fréquentes dans les microblogs et juge l'intérêt des utilisateurs. Elle permet d'identifier les sujets les plus émergents au sein des réseaux sociaux ainsi de surveiller un événement bien spécifique dans les flux des microblogs, à ce fait on peut déclencher des alertes ou bien des avertissements (en cas de séisme / épidémie).

- ***Recherche de microbloggeurs***

La recherche de microbloggeurs repose sur des objectifs tels que l'identification des utilisateurs les plus populaires autrement dit : ceux qui ont les mêmes centres d'intérêts que l'utilisateur courant, ou bien les experts dans des domaines particuliers. Plusieurs travaux [9] se sont focalisés sur l'identification des utilisateurs les plus populaires dans les plateformes de microblogging car ces derniers jouent un rôle clé au niveau du réseau social.

II.2.4.2 Critères de pertinences de Twitter

Le but est de restituer les microblogs contenant la réponse qui satisfait le besoin en information de l'utilisateur par la recherche de microblog. Pour évaluer la pertinence de celui-là, on prend en compte de divers critères. Parmi ces derniers il y a ceux qui sont liés au réseau social comme le facteur temporel ainsi que la popularité de l'auteur. Parmi ces critères :

- ✓ ***Le critère de pertinence thématique***

Vu que la thématique assure la similarité entre le microblog et la requête, afin de l'estimer (Robertson, 2004 [16]) a utilisé la fréquence des termes dans son algorithme de pondération

(TF*IDF).Mais les modèles de la RI classique sont limités à cause de la faible longueur du microblog, d'où est venue l'idée de l'expansion des requêtes ou bien l'enrichissement des microblogs avec des termes de micoblogs similaires.

- ✓ ***Le critère de pertinence temporelle***

En général, l'utilisateur a un besoin d'accès directe à des informations récentes telles que les événements et les buzzs, ce qui a motivé la recherche de microblogs. De ce fait, les

tweets les plus récents ont plus de chance d'être pertinents par rapport à des tweets publiés antérieurement.

Pour cela, à ce niveau, le « score de fraîcheur du document » est calculé en termes de différence temporelle entre la date de la soumission de la requête et la date de publication du document.

- ✓ ***Critère de pertinence sociale***

Ce critère décrit l'importance sociale des utilisateurs ainsi celle des microblogs à travers des métriques et des mesures. Une bonne recherche dans les microblogs ne doit présenter que des ressources fiables autrement dit la pertinence est liée à la crédibilité de la source d'information.

Selon l'approche (Duan et al 2010), un tweet est jugé pertinent selon le nombre de mentions, le nombre de fois qu'il est retweeté mais aussi l'auteur qui l'a tweeté , il doit être important au sein du réseau social.

- ✓ ***Critère de localisation géographique***

La localisation géographique d'un tweet joue un rôle dans l'évaluation de la pertinence de ce dernier. En effet, certains utilisateurs recherchent des tweets à propos d'un sujet propre à une région particulière de ce fait, un tweet publié à l'endroit où est produit l'événement en question et plus pertinent que celui publié dans un lieu différent que celui où est produit l'événement. En résumé, les tweets géographiquement proches de leur localisation ont plus tendance à intéresser les utilisateurs.

- ✓ ***Critère de pertinence: sentiments***

Un microblog reflétant des sentiments sur une personnalité, des événements ou un produit en particulier est pertinent lors d'une recherche des avis ou opinions sur ces derniers.

- ✓ ***Critère de pertinence : Longueur du microblog*** :

La Longueur du microblog est le nombre de termes dans celui-là. Selon (Zhao et al [14], 2011) La longueur d'une phrase reflète la quantité d'information qu'elle véhicule.

- ✓ ***Fréquence de retweets*** :

Autrement dit nombre de fois qu'un tweet a été retweeté. D'après (Zhao et al., 2011 ; Magnani et al.,2012 ; Vosecky et al., 2012 [10] ; Duan et al., 2010 [12])Si un tweet est repartagé, ça veut dire que son contenu intéressant ce qui implique qu'il est pertinent.

- ✓ ***Fréquence de hashtags*** :

Représente le nombre de hashtags dans un tweet. (Duan et al., 2010)ont confirmé qu'ils sont utilisés pour définir un topic pour le tweet, ou bien pour s'intégrer à une conversation.

- ✓ ***Réponse*** :

Ce critère montre que le tweet ne s'agit pas d'un message isolé et sans interaction grâce aux réponses des autres microblogs (Vosecky et al., 2012 [10] ; Metzler et Cai, 2011 [11] ; Duan et al., 2010 [12]).

II.2.5 L'évaluation de la RI dans Twitter

Comme en recherche d'information classique, l'évaluation se fait à travers les collections de tests dans le but de mettre en oeuvre les approches de restitution de microblogs pertinents. Ces collections sont souvent construites dans le cadre de campagnes d'évaluation. Parmi ces dernières on en cite la plus populaire surnommée TREC.

• La tâche TREC Microblog

C'est une tache de la compagne d'évaluation TREC dédiée à la RI dans les microblogs, elle évalue les méthodes et les approches de RI dans les plateformes de microbloggings (Twitter).

Elle est décrite comme une tache exprimée sous forme de mots clés (tache ad hoc). Cependant, les systèmes doivent répondre à la requête par une liste de documents classés par ordre de pertinence par rapport au moment de soumission de la requête. Par conséquent, l'information la plus récente et la plus pertinente sera retournée à la demande de l'utilisateur.

La première compagne TREC Microblog fournit le corpus Tweets2011, qui comprend:

**Environ 16 millions de tweets qui ont été publiés sur période approximative de deux semaines. Le corpus est considéré comme un échantillon fiable de la twittosphère.

**Et 50 requêtes (topics) dont chacune représente un besoin en information à un moment donné ci-suit :

un aperçu du format des topics.

<top>

<num>Number : MB007 </num>

<title>Pakistan diplomat arrest murder</tilte>

<querytime> Tue Feb 08 22 :56:33 +0000 2011 </querytime>

<querytweettime>35109758973255680 </querytweettime>

</top>

Conclusion

Dans la première partie de ce chapitre, nous avons présenté l'information sociale, développée avec l'évolution des technologies du Web 2.0. Nous avons ensuite décrit les types d'informations sociales sur internet, en particulier, le contenu généré par l'utilisateur et celui généré par la pratique. Par la suite, nous avons discuté l'impact de l'évolution de ces informations sociales sur le processus de RI, ainsi que leur emploi dans le but d'améliorer l'efficacité des SRI, enfin nous avons abordé les approches qui les exploitent en fonction du niveau de leur utilisation soit du côté utilisateur ou bien du côté document.

Dans la deuxième, nous avons abordé la RI dans les microblogs, nous avons parlé de Twitter, une plateforme de microblogging. Et avons vu les concepts de son fonctionnement ainsi ses caractéristiques. Puis nous nous sommes approfondies dans la RI dans Twitter en nous intéressant aux approches de recherches dans celui-ci. De plus nous avons exploré les critères de pertinences de ces derniers. Ensuite, nous avons cité et présenté l'une des compagnes d'évaluation de la RI dans les microblogs TREC Microblog.

Chapitre III

Etat de l’art de la recherche d’information dans Twitter

Chapitre III

Etat de l’art de la recherche d’information dans Twitter

Introduction

Depuis quelques années, la quantité d'information publiée sur les plateformes de microblogging augmente d'une manière exponentielle. Prenons par exemple le cas de Twitter où le nombre de publications peut atteindre une centaine de milliers de microblogs ou tweets. De ce fait, plusieurs experts de domaine se sont intéressés pour remédier au problème d'extraction d'informations pertinentes des microblogs.

III.1 Intégration des signaux sociaux dans la RIS

Les signaux sociaux sont des informations qui fournissent des renseignements sur les interactions, les émotions, les relations et les comportements sociaux d'un utilisateur à travers des fonctionnalités offertes par les RS, tels que le nombre de partages, de commentaires, de tags, les j'aime, les retweets …etc.

Pour ce faire, nous allons aborder les approches déjà proposées qui intègrent les signaux sociaux dans la RI. Nous en citons :

Approches de [SEKOUR.M] : il a proposé deux approches. Dans la première il utilise les signaux sociaux afin d'améliorer les résultats de recherche en les ajoutant à l'étape d'indexation pour favoriser les tweets les plus populaires.

Pour ce faire , il propose un triplet (U, M_S, T)

Où :

- U : est l'ensemble utilisateur
- M_S : est l'ensemble des métadonnées contenant le nombre de followers N_f lenombre de commentaires N_c le nombre de retweets N_r
- T : l'ensemble de tweets

Cette approche consiste à modifier le processus d'indexation des tweets et cela en prenant en compte les signaux sociaux des tweets en plus de leurs contenus textuels dans le but de différencier entre eux par leur popularité. Autrement dit, ceux possédant un taux de retweets élevé.

Et la modification se porte au niveau de l'indexation dans la fonction de pondération, dans le but de prendre en considération les métadonnées (signaux sociaux) relatives à chaque tweet .

$$\omega = \alpha (TF \cdot IDF) + (1 - \alpha) \cdot P(t) \quad \textbf{(3.18)}$$

SEKOUR dans cette approche a calculé le Tf-Idf standard auquel il a rajouté une certaine pertinence sociale propre à chaque tweet***P(t)***.

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

Où il donne l'importance à la pondération textuelle et aussi favorise les retweets par rapport aux autres signaux du tweets en question.

Où :

TF : fréquence du terme dans le *Tweet* (TermFrequency)$TF_{ij} = \frac{f(t_i,T_j)}{\sum_k (t_i,T_j)}$

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'aimé ».

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 $1<\gamma<3$..

Dans la seconde approche il utilise les informations de twitter comme ressource externe pour l'expansion de requête afin de bien cerner le besoin en information de l'utilisateur, et cela consiste d'abord à faire une recherche dans la collection de tweets avec une requête passée par l'utilisateur en utilisant sa première approche afin d'indexer les tweets en exploitant les signaux sociaux issus de chaque tweet, il aura comme sorties les documents pertinents.

Vu que son but est de rapprocher le plus possible la requête vers les tweets pertinents . En vue d'une bonne reformulation de la requête, il propose une variante du pseudo relevance feedback de Rocchio. [17]

Approche de [HANNACHI.F] : dans son approche HANNACHI combine les relations d'abonnement et de retweets dans un score d'influence unique. En vue d'une bonne amélioration de la RI dans Twitter, l'intuition derrière l'utilisation de ces deux signaux sociaux vient du faitque :

- Si un microblogger est influent alors le nombre de ses abonnés est relativement élevé par rapport à ses abonnements.
- De plus, si un microblogger est plus « suivi » que « suiveur » alors il est très probablement un blogueur influent.
- Si un twitto est influent alors le nombre de retweets liés à ses tweets est

Probablement, considérablement élevé.

- De même, si les tweets d'un blogueursont fortement retweetés, ce blogueur devient par ce fait influent.

De ces observations lui est venue l'idée de proposer les mesures baséessur les relations d'abonnement, retweet et de mention. Pour cela elle définit deux paramètres de mesure de l'influence :

- **Le ratio d'abonnement :** est le rapport entre le nombre d'abonnés et le nombre d'abonnements
- **Le ratio de retweet :** définit comme la somme des nombres de retweets et de mentions divisé par le nombre de tweets publiés

Elle propose de combiner les deux ratios précédents en un score additif pondéré. [18]

Approche de [BENJABEUR] : [19] cette approche combine la pertinence thématique et l'importance sociale des blogueurs.

Ce modèle considère l'influence et l'expertise comme les principaux facteurs sociaux qui déterminent l'importance du blogueur et la qualité de ses articles, l'influence d'unblogueur dépend de ses relations de retweets étant estimées selon sa position dans le réseau social d'influence. Quant à l'expertise, celle-ci est déterminée par la distribution des termes dans ses articles suivant un modèle de langue.

- ➔ Selon BENJABEUR, quand un utilisateur retweet un article, il confirme l'importance du message communiqué et exprime son intérêt au sujet et adopte la même idée si une opinion y est exprimée.
- ➔ Il a constaté que les abonnés continuent à rediffuser les messages s'ils jugent leur contenu est important.
- ➔ L'influence d'un blogueur est alors déterminée par la proportion de ses messages rediffusés.

Approche de [SAVONNET. M et FRAME.A]:[20] cette approche repose sur une méthode de mesure d'influence des candidats sur twitter. Selon ses auteurs, leur modèle reconnaît trois actions qui peuvent se produire : la réponse, le retweet, la mention, le hashtag , le lien URL, le favoris et le nombre de followers, destinées à propager le contenu à d'autres utilisateurs et sont la preuve et marqueurs de l'influence qui s'est produite.

III.2 La temporalité dans la RI

Dans le but d'améliorer les modèles de la RI classique la RI temporelle exploite les Informations temporelles pouvant exister dans les documents et les requêtes, et cela en combinant de plus que la pertinence thématique, la dimension temporelle dans le but est de restituer les documents récents et de répondre aux requêtes des utilisateurs (pertinence sociale).

Le temps est représenté par la date de création des documents avec la date de soumission d'une requête comme l'approche de MASAKI AONO, 2015) ou par les expressions temporelles contenues dans les documents comme notre proposition. Ace niveau nous allons citer quelques approches exploitant la temporalité :

Approche de [DJEDDI.A et BENDOU.A] [21] : ils ont proposé deux approches pour la recherche sociale des tweets qui associent la pertinence thématique à l'importance sociale des tweets correspondants (pertinence sociale). Ils tentent d'exploiter le nombre de retweets, identifier leurs vitesses dans le but de déterminer l'importance sociale d'un tweet.

Selon cette approche, l'importance du tweet retransmis traduit l'importance du message véhiculé.

Du coup l'importance d'un tweet est alors déterminée par son nombre de retweets dans une certaine période de temps.

En vue de restituer un tas de tweets qui seront à la hauteur de couvrir le sujet de la requête, ils combinent un score de pertinence thématique et un score de pertinence sociale du tweet possédant une vitesse de retransmission plus grande.

Ces deux scores sont combinés linéairement ainsi :

$$Score(Q, T_i) = \alpha \cdot Score_{Thématique}(Q, T_i) + \beta \cdot Score_{Social}(T_i, D, Nbre\_RT) \qquad (3.19)$$

Où :
- ***Q :*** est la requête
- T_i est le Tweet
- ***D :*** est le temps pris par le tweet pour atteindre un certain nombre de retweets
- ***Nbr_RT*** : le nombre de retweets.
- ***α et β*** sont des coefficients d'amortissements, obtenus par expérimentations afin d'optimiser les résultats

⇨ Pour la pertinence thématique : ils ont utilisé le modèle vectoriel de LUCENE.

⇨ Pour le score social : il évalue l'importance d'un tweet T_i en fonction de son nombre de retweets ainsi :

$$Score_{Social}(T_i, D, Nbr\_RT) = \mu_{Nbr\_RT} + Score_{Vitesse\_RT}(T_i, D, Nbr\_RT) \qquad (3.20)$$

Cette approche privilégie les tweets ayant un taux élevé de retweets avec μ_{Nbr_RT}

Où :
- $\mu_{Nbr\_RT} = 0,5$ **si** $Nbr\_RT > 50$
- 0 sinon

Avec $Score_{Vitesse\_RT}(T_i, D, Nbr\_RT)$ ils calculent la vitesse de retweet du Tweet

T_i ce Score est le rapport entre le nombre de retweets ***Nbr_RT*** et le temps qu'il fallut au Tweet pour l'atteindre. Il est calculé ainsi :

$$Score_{VitesseRT}(T_i, D, Nbr\_RT) = \frac{Nbr\_RT}{D} \qquad (3.21)$$

Tel que:

D = Date du dernier retweet du Tweet T_i – Date de création du tweet T_i

Vu que l'approche 1 n'a pas donné de résultats satisfaisants. Ils ont proposé une autre manière de calcul de la vitesse des retweets ainsi :

- Calculer la vitesse du premier retweet
- Calculer la vitesse du deuxième retweet, puis du troisième, jusqu'au dernier.
- Calculer la somme des vitesses de tous les retweets.
- Diviser la somme des vitesses par le nombre retweets.

Ce procédé prend en compte toute la durée des retweets, de la date du 1er retweet à la date du dernier retweet, ce qui garantit un traitement égal pour les tweets influents sur une longue durée. Ce qui évite les passages à vide.

➔ Pour la thématique et la pertinence sociale sont combinées linéairement de la même manière que qu'en approche 1.

➔ Pour le calcul du score social **:** le score social $Score_{Social}\ (T_i\ ,\ D,\ Nbr\_RT)$ évalue l'importance d'un tweet ***Ti*** en fonction de son nombre de retweets ***Nbre_RT*** la durée ***D*** de retweet numéro ***j*** ainsi que son rang R_i

$$Scores_{ocial}(\ T_i, D, Nbr\_RT\ ) = Score_{Vitesse\_RT}\ (T_i\ ,\ R_{ij}D_{ij}\ ,\ Nbr\_RT_{ij}) \qquad \textbf{(3.22)}$$

Où :

- $T_i$le tweet, R_{ij} le retweet de T_i numéro ***j***
- ***Dij***la durée du retweet numéro ***j*** du tweet T_i
- Nbr_RT_{ij} le classement du retweet donc le nombre de retweets à l' instant ***j***.

$$ScoreVitesse\_RT\ (Ti, Rij\ , Dij, Nbr\_RTij) = \sum \frac{\frac{Nbr\_RTij}{Dij}}{Nbr\_RT} \qquad \textbf{(3.23)}$$

Et :

D_{ij}= Date du retweet numéro j du Tweet T_i – Date de création du tweet T_i

Approches de [DAMAK] :il propose trois approches différentes qui intègrent le temps:

- Il propose d'augmenter le score social d'un tweet en fonction de sa proximitétemporelle avec la date de la requête.
- En outre, il favorise les termes fréquemment utilisés au moment du rafale.

L'emploi de la fraîcheur dans les deux méthodes proposées n'apporte pas d'amélioration.

- Dans cette troisième méthode, DAMAK propose d'amplifier le score d'un terme dans un tweet publié à un instant *t* en fonction de la fréquence d'emploi de ce terme dans cette période *t*. Un même terme aura des scores différents en fonction de la date de soumission du document auquel il appartient.

Ce score sera plus important si le terme appartient à un document publié dans une période de rafale de ce terme, que dans le cas où il appartient à un document publié dans une période où le terme n'est pas fréquemment utilisé.[8]

La prise en compte de la fraîcheur de cette façon n'a pas montré aussi son effet.

Approche de[MASAKI AONO] : cette approche propose un reclassement de résultats obtenus lors d'une recherche effectuée avec un modèle de RI classique en extrayant les différents critères du tweet puis les combinant avec le score d'appariement requête/document et cela afin d'estimer la pertinence. [23]

MASAKI exploite le temps comme facteur mesurant la proximité temporelle entre la date de publication du tweet $\boldsymbol{Tweet_{Time}}$ et la date de soumission de la requête $\boldsymbol{Querry_{Time}}$, comme suit **:**

$$\boldsymbol{TimeScore} = \frac{\mathbf{1}}{\sqrt{\boldsymbol{Querry_{Time} - Tweet_{Time}} + \mathbf{1}}} \qquad \textbf{(3.24)}$$

Approche de[Willis] : cette approche intègre la temporalité pour améliorer la recherche dans les microblogs**.** Les auteurs de cette approche proposent l'expansion de requête basée sur le temps, quis'intéresse à la priorité temporelle,favorisation des termes qui ont une haute occurrence avec tous les termes de larequête: [24]

• Utilisation de la technique du regroupement pour marquer les termes candidats d'expansion basée sur la récence, et cela :

**Lors de la recherche initiale, en identifiant les périodes de temps pertinentes,

**Puis en regroupant les meilleurs résultats par date/ heure.

**Les groupes sont classés selon la taille par ordre décroissant et indexé ainsi :

i= {1,…, T}.

**Le premier groupe (i=1) correspond à celui qui a le plus grand nombre de tweetsdes n premiers résultats, le dernier groupe (i=T) correspond à celui avec le plus petit nombre de tweets des n premiers résultats,

**Puis, on favorise les résultats des plus grands groupes (c'est-à-dire période associée aux meilleurs résultats).

Conclusion

Dans ce chapitre, nous avons passé en revue la notion de la temporalité dans la recherche d'information dans Twitter. Nous avons introduit des approches distinctes qui se basent sur différentes informations disponibles sur la plateforme de microblogging, chacune de ces approches a su montrer des résultats plus ou moins satisfaisants, dans le chapitre suivant nous détaillerons l'approche que nous avons proposée.

Chapitre IV

Proposition et expérimentation de l’approche

Introduction

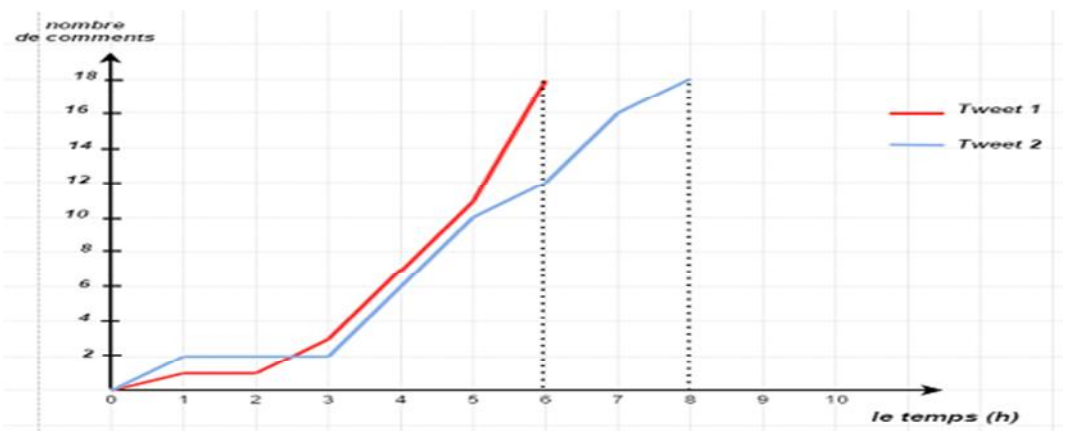
Dans le chapitre précédant, nous avons présenté un état de l'art de la RI dans la plateforme de microblogging Twitter, où nous avons vu quelques approches qui intègrent les signaux sociaux ainsi que le facteur temps afin d'améliorer la RI dans Twitter. Dans ce qui suit nous proposons une approche basée sur la pertinence thématique et la pertinence sociale du tweet qui exploite les commentaires pendant une certaine durée, sachant que nous avons suivi les perspectives proposées par Mr DJEDDI.A et BENDOU.A.

A. Proposition de l'approche

1. Principe

L'importance du tweet transmis est confirmée par son nombre de commentaires .Plus le tweet possède un nombre de commentaires élevé plus il est pertinent.

Dans la figure ci-dessous nous avons estimé le nombre de commentaires dans une période de temps



Figure(9) : - Exemple qui illustre la vitesse de deux tweets -

La figure nous montre deux tweets ayant 18 commentaires chacun. Le $Tweet_1$ a réussi à avoir les 18 commentaires en seulement 6h, or que $Tweet_2$ a pris 8h. Peut-on conclure que le $Tweet_1$ est plus pertinent que $Tweet_2$?

Dans ce qui suit nous proposons une approche basée sur le nombre de commentaires d'un tweet en fonction du temps.

2. Approche proposée

Notre approche combine deux scores: un score de pertinence thématique et un score de pertinence sociale du tweet, afin de répondre aux besoins en information de l'utilisateur.

Le score thématique dépend du tweet et de la requête, nous l'avons calculé en utilisant le modèle vectoriel de LUCENE. Pour le score social qui détermine l'importance d'un tweet , nous avons calculé la vitesse de ses commentaires dans le temps.

Sa formulation est comme suit:

$$Score\ (Q, Ti) = \alpha * ScoreThematique\ (Q, Ti) + \beta * ScoreSocial\ (T_i) \quad \textbf{(4.25)}$$

:

Où :

- ***Q :*** est la requête
- T_i **:** est le tweet.
- ***α et β :***sont des coefficients d'amortissement calculés expérimentalement, leur rôle est d'ajuster le poids de l'équation ; tel que ***α et β*** sont dans l'intervalle]0.1[et $\alpha+\beta=1$

- ***Le calcul de la pertinence sociale « $Score_{Social}$ »*** il s'agit de calculer la vitesse des commentaires de chaque tweet.

$$ScoreSocial\,(T_i) = \frac{\sum \frac{NC_{ij}}{D_{ij}}}{NC_i} \quad \textbf{(4.26)}$$

Où :

- ***Ti*** le tweet,
- NC_{ij}le commentaire de ***Ti*** numéro ***j***,
- D_{ij} la durée du commentaire numéro ***j*** du tweet ***Ti***
- NC_ile nombre de commentaires du tweet T_ià l' instant ***j***.

La durée ***D*** est calculée ainsi :

$$D_{ij} = Date\ du\ j^{ème}\ commentaire\ du\ Tweet\ T_i - Date\ de\ création\ du\ tweet\ T_i \quad \textbf{(4.27)}$$

Nous avons constaté que si une requête ne possède aucun tweet pertinent, c'est à dire le score thématique est égale à 0. Toutefois, le score social est calculé sur tous les tweets et seront réordonnés par le SRI malgré que la thématique est égale à 0. Cependant, des tweets

non pertinents seront retournés; pour y remédier nous avons proposé une autre formulation pour notre approche:

$$Score\ (Q, T_i) = \alpha * ScoreThematique\ (Q, T_i) + \beta * ScoreThematique\ (Q, T_i) * ScoreSocial\ (T_i)$$

$$Score\ (Q, Ti) = ScoreThematique\ (Q, T_i) \square\ (\alpha + \beta * ScoreSocial\ (T_i)) \quad \textbf{(4.28)}$$

Où :

- ***Q:*** est la requête
- ***Ti :*** est le tweet.
- ***α et β :***sont des coefficients d'amortissement calculés expérimentalement, leur rôle est d'ajuster le poids de l'équation ; tel que ***α et β*** sont dans l'intervalle]0.1[et***α+β*=1**

ScoreThematique (Q, Ti) et ***ScoreSocial (Ti ,)*** sont calculés comme précédemment .

B. Evaluation et expérimentation

Nous avons présenté notre approche qui intègre le facteur temps dans la recherche des microblogs, et cela en utilisant les commentaires des tweets pendant une certaine durée. Dans ce qui suit, nous allons présenter les outils qui ont contribués à la réalisation de notre application. Puis, décrire le cadre expérimental de notre proposition, en outre, nous allons présenter les résultats obtenus et les discuter, en les comparant avec ceux de la thématique de LUCENE.

1. Outils de développement

i. Eclipse IDE

Eclipse est un environnement de développement (IDE) placé en open source. En plus de java, Eclipse peut être utilisé avec d'autres langages de programmation tel que PHP et C/C++. La spécificité d'Eclipse IDE vient du fait que son architecture totalement développée autour de la notion de plug-in .Il est gratuit et disponible pour la plupart des systèmes d'exploitation qu'on peut trouver sur le marché.

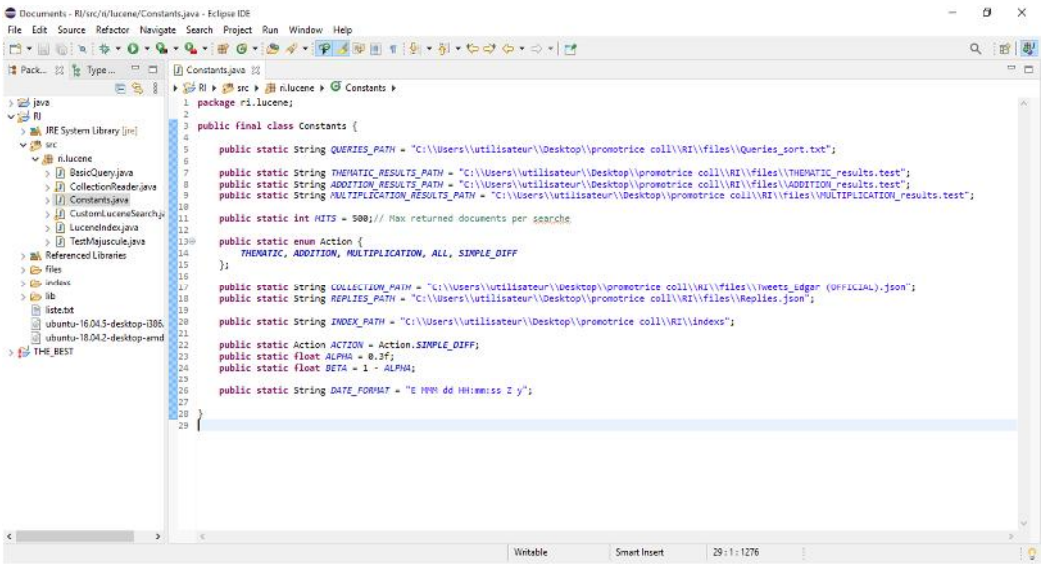


Figure (10) : - Aperçu de l'interface de l'IDE Eclipse. -

ii. Le langage Java

Nous avons utilisé le langage java, qui est un langage de programmation orienté objet créé par James Gosling, Patrick Naughton et Bill Joy, présenté officiellement le 23 mai 1995 au «***SunWorld***» une société rachetée en 2009 par la société Oracle qui détient Java.

Les logiciels écrits dans java sont compilés vers une représentation binaire intermédiaire qui peut être exécutée dans une machine virtuelle Java (JVM), cela par l'abstraction du système d'exploitation.

La particularité et l'objectif central de Java est que les logiciels écrits dans ce langage doivent être très facilement portables sur plusieurs systèmes d'exploitation tels que UNIX, Windows, Mac Os ou GNU/Linux, avec peu ou pas de modifications. Pour cela, divers plateformes et Framework associés visent à guider, sinon garantir, cette portabilité des applications développées en Java est assurée par ce langage.

iii. Lucene

Lucene est une bibliothèque open source écrite en Java qui permet l'indexation et la recherche du texte. Il est utilisé dans certains moteurs de recherche. C'est un projet de la fondation Apache étant mis à disposition sous licence Apache. Il est également disponible pour les langages Ruby, Perl, C++, PHP, C#, Python.

iv. Description de la collection de tests

Nous avons utilisé la collection Tweets_Edgar, une collection gratuite contenant : 502 tweets publiés entre mai 2010 et mars 2011, 802 requêtes et 802 jugements de pertinence. Mais celle-ci ne répondait pas à nos besoins vu qu'elle ne contient pas suffisamment de commentaires, et ceux présents ne répondent pas aux tweets de la collection. Vu que nous n'avons pas pu récupérer les commentaires des tweets que nous possédons au niveau de l'API Twitter4J nous avons opté pour une solution, qui est de récupérer les tweets de la collection puis de générer un fichier aléatoire où nous avons affecté aux tweets extraits de la collection un nombre de commentaires qui ne dépasse pas 50 commentaires.

→ Nous mettons en évidence les quatre classes principales suivantes :

LuceneIndex: responsable de l'indexation

CustomLuceneSearch: responsable de génération des états de sortie en quatre modes d'exécution (ACTION) :

- ✓ **THEMATIC** : génère la sortie sur le fichier THEMATIC_resuls.test
- ✓ **ADDITION** : génère la sortie sur le fichier ADDITION_resuls.test
- ✓ **MULTIPLICATION** : génère la sortie sur le fichier MULTIPLICATION_resuls.test
- ✓ **ALL** : génère la sortie des trois actions précédentes à la fois.

CollectionReader :

- Lire les fichiers de collection [Tweets_Edgar (OFFICIAL).json, Replies.json] et les convertir en documents indexables par Lucene et associer à chaque document son score social après l'avoir calculé.
- Génère une collection de replies fictifs basés sur la date de création des tweets et leurs IDs

Constants: paramétrer le programme à savoir :

- Les chemins vers le fichier de Requêtes, Collection, commentaires, Resultat (Thématique, Addition, Multiplication) et le dossier d'index
- Les Variables ALPHA, BETA
- Mode d'exécution de CustomLuceneSearche ou l'action dont nous avons parlé juste avant

2. Outil d'évaluation

➢ Trec-Eval

C'est un outil utilisé afin d'évaluer l'ordre des documents triés par pertinence. Pour cela, on utilise les deux fichiers Qrels et Results définis ainsi :

- Qrels : répertorie les jugements de pertinence pour chaque requête
- Résults : contient le classement des résultats renvoyés par le SRI.

Trec_eval est open source, après l'avoir téléchargé on y accède via des commandes tapées sur Ubuntu.

Pour pouvoir exécuter Trec_eval, il suffit de taper la commande:

$ ./trec_eval -q -c qrels_file result_file

➢ Mesures d'évaluation

L'évaluation de notre travail se base sur les mesures suivantes :

- La MAP
- La R_Precision
- La Precision@X
- La précision moyenne
- Le rappel, la précision et la F-mesure
- Le rappel interpolé

3. Résultats des scores

Voici quelques résultats des scores obtenus suite à la recherche thématique respectivement à celle de la formule I :

$$Score\ (Q, Ti) = \alpha * ScoreThematique\ (Q, Ti) + \beta * ScoreSocial\ (T_i) \quad \textbf{(4.25)}$$

$/\ \alpha + \beta = 1 et\ \boldsymbol{\alpha = 0.6}$

Et à celle de formule II

$$Score\ (Q, Ti) = \alpha * ScoreThematique\ (Q, T_i) + \beta * ScoreThematique\ (Q, T_i) * ScoreSocial\ (T_i)$$

effectuées sur les requêtes de notre collection Tweets_Edgar :

10105828	Q0	48590675775725568	1	9,7238817215	STANDARD
10105828	Q0	49208114750300160	2	4,7551627159	STANDARD
10105828	Q0	48533257020641280	3	4,6075391769	STANDARD
10106	Q0	49751557398462464	1	6,3940691948	STANDARD
10106	Q0	46184722304483328	2	5,8868331909	STANDARD
10106	Q0	46239714512093184	3	4,3381824493	STANDARD
10244706	Q0	48573622926843904	1	7,3431482315	STANDARD
10244706	Q0	49222663318679552	2	5,4541583061	STANDARD
10244706	Q0	48473709756952576	3	4,4688057899	STANDARD
103067	Q0	44590727145066497	1	7,6300110817	STANDARD
103067	Q0	18815381573472256	2	7,3205857277	STANDARD
103067	Q0	45506571496730624	3	4,9125590324	STANDARD
103067	Q0	43697536430637056	4	4,0582408905	STANDARD
103067	Q0	49438120193699840	5	3,5590269566	STANDARD
103067	Q0	48095255081394176	6	3,4365282059	STANDARD
103067	Q0	49685777046716416	7	3,4365282059	STANDARD
103067	Q0	33781064761614336	8	3,180560112	STANDARD
103067	Q0	43429826794766336	9	3,180560112	STANDARD
103067	Q0	49883233193705472	10	3,0716791153	STANDARD
103067	Q0	49291616581722112	11	3,0716791153	STANDARD
103067	Q0	50314953617440768	12	3,0716791153	STANDARD
103067	Q0	17959296583077888	13	2,9700062275	STANDARD
103067	Q0	49748582445744129	14	2,9700062275	STANDARD
103067	Q0	45323108768419840	15	2,7855992317	STANDARD
103067	Q0	48473709756952576	16	2,7017245293	STANDARD
103067	Q0	43658543689244672	17	2,6227529049	STANDARD
103067	Q0	49259775644549121	18	2,6227529049	STANDARD
103067	Q0	43783485965602816	19	2,5482668877	STANDARD
103067	Q0	48465971043778560	20	2,5482668877	STANDARD
103067	Q0	50071335975653376	21	2,5482668877	STANDARD
103067	Q0	45911529283010560	22	2,4778950214	STANDARD
103067	Q0	49869747491323904	23	2,4778950214	STANDARD
103067	Q0	48448177854103553	24	2,348200798	STANDARD
10396687	Q0	49759546402545665	1	6,3940691948	STANDARD
10396687	Q0	47650806131994625	2	5,6622419357	STANDARD
10396687	Q0	49864149886451713	3	5,5918159485	STANDARD

Figure (11) : - Aperçu des résultats du score thématique.-

10105828	Q0	48590675775725568	1	7,3035216331	STANDARD
10105828	Q0	49208114750300160	2	3,9361743927	STANDARD
10105828	Q0	48533257020641280	3	3,9337496758	STANDARD
10106	Q0	46184722304483328	1	5,6127271652	STANDARD
10106	Q0	49751557398462464	2	4,4216270447	STANDARD
10106	Q0	46239714512093184	3	3,3531742096	STANDARD
10244706	Q0	48573622926843904	1	5,0066947937	STANDARD
10244706	Q0	49222663318679552	2	3,6363046169	STANDARD
10244706	Q0	48473709756952576	3	3,4012835026	STANDARD
103067	Q0	44590727145066497	1	5,8173398972	STANDARD
103067	Q0	18815381573472256	2	4,5594944954	STANDARD
103067	Q0	48095255081394176	3	3,6761698723	STANDARD
103067	Q0	49291616581722112	4	3,6714882851	STANDARD
103067	Q0	43697536430637056	5	3,5349049568	STANDARD
103067	Q0	49869747491323904	6	3,5252289772	STANDARD
103067	Q0	49883233193705472	7	3,3181056976	STANDARD
103067	Q0	45506571496730624	8	3,1642022133	STANDARD
103067	Q0	43429826794766336	9	3,0996286869	STANDARD
103067	Q0	48465971043778560	10	2,7759795189	STANDARD
103067	Q0	49438120193699840	11	2,772559166	STANDARD
103067	Q0	17959296583077888	12	2,7691464424	STANDARD
103067	Q0	48448177854103553	13	2,7626347542	STANDARD
103067	Q0	43783485965602816	14	2,5736031532	STANDARD
103067	Q0	49748582445744129	15	2,5102577209	STANDARD
103067	Q0	49685777046716416	16	2,3857264519	STANDARD
103067	Q0	48473709756952576	17	2,3410346508	STANDARD
103067	Q0	50071335975653376	18	2,3076415062	STANDARD
103067	Q0	50314953617440768	19	2,0985631943	STANDARD
103067	Q0	33781064761614336	20	1,9083361626	STANDARD
103067	Q0	45323108768419840	21	1,8713595867	STANDARD
103067	Q0	43658543689244672	22	1,8236517906	STANDARD
103067	Q0	45911529283010560	23	1,7022131681	STANDARD
103067	Q0	49259775644549121	24	1,6936517954	STANDARD
10396687	Q0	49759546402545665	1	6,1381511688	STANDARD
10396687	Q0	49864149886451713	2	5,2497324944	STANDARD
10396687	Q0	47650806131994625	3	3,9688472748	STANDARD

Figure (12) : - Aperçu des résultats du score de la formule I . -

10105828	Q0	48590675775725568	1	20,1205825806	STANDARD
10105828	Q0	48533257020641280	2	8,1517782211	STANDARD
10105828	Q0	49208114750300160	3	8,0033044815	STANDARD
10106	Q0	46184722304483328	1	15,7804059982	STANDARD
10106	Q0	49751557398462464	2	7,5781559944	STANDARD
10106	Q0	46239714512093184	3	5,857694149	STANDARD
10244706	Q0	48573622926843904	1	8,8176956177	STANDARD
10244706	Q0	48473709756952576	2	5,8988232613	STANDARD
10244706	Q0	49222663318679552	3	5,2567696571	STANDARD
103067	Q0	44590727145066497	1	14,0341329575	STANDARD
103067	Q0	48095255081394176	2	7,6093425751	STANDARD
103067	Q0	49291616581722112	3	7,4595131874	STANDARD
103067	Q0	43697536430637056	4	6,8988480568	STANDARD
103067	Q0	49869747491323904	5	6,5379066467	STANDARD
103067	Q0	49883233193705472	6	6,3740353584	STANDARD
103067	Q0	43429826794766336	7	5,6973137856	STANDARD
103067	Q0	18815381573472256	8	5,6159353256	STANDARD
103067	Q0	17959296583077888	9	4,7138237953	STANDARD
103067	Q0	48465971043778560	10	4,7066979408	STANDARD
103067	Q0	48448177854103553	11	4,5877132416	STANDARD
103067	Q0	49438120193699840	12	4,4030246735	STANDARD
103067	Q0	43783485965602816	13	4,1909890175	STANDARD
103067	Q0	45506571496730624	14	4,0119233131	STANDARD
103067	Q0	49748582445744129	15	3,9449226856	STANDARD
103067	Q0	48473709756952576	16	3,5662763119	STANDARD
103067	Q0	50071335975653376	17	3,5132482052	STANDARD
103067	Q0	49685777046716416	18	3,1746973991	STANDARD
103067	Q0	50314953617440768	19	2,6279921532	STANDARD
103067	Q0	43658543689244672	20	2,2293400764	STANDARD
103067	Q0	45323108768419840	21	2,2284793854	STANDARD
103067	Q0	45911529283010560	22	2,0206644535	STANDARD
103067	Q0	33781064761614336	23	1,9083361626	STANDARD
103067	Q0	49259775644549121	24	1,8883821964	STANDARD
10396687	Q0	49759546402545665	1	18,5537319183	STANDARD
10396687	Q0	49864149886451713	2	13,9495830536	STANDARD
10396687	Q0	13848368514	3	8,447303772	STANDARD

Figure (13) :- Aperçu des résultats du score de la formule II . –

4. Résultats des mesures d'évaluation et discussion

Formule I :

- **La precision@X**

	P@5	**P@10**	**P@20**
Thématique	0.1693	0.0919	0.0477
formule I	0.1703	0.0920	0.0477

Tableau (1) : P@5, P@10, P@20 pour la thématique et la formule 1.

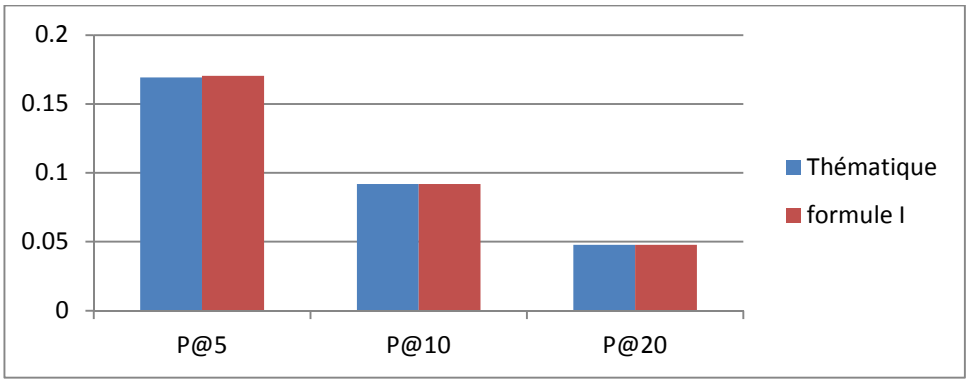


Figure (14) : - Comparaison des Précision@X du score de la formule I et de celui de la thématique-

Selon ces résultats, nous remarquons que chacun des deux scores thématiques et celui de la formule I perdent leurs précisions à chaque augmentation des documents restitués, et nous avons une amélioration dans la précision@X au niveau de P@5 avec 0.0193 et aussi au niveau de P@10 une amélioration de 0,0001 par rapport au score thématique.

- **La R-précision, la MAP et la précision moyenne**

	R-précision	MAP	Précision moyenne
La thématique	**0,5639**	**0,6018**	**0,0282**
L'approche I	**0,5547**	**0,5963**	**0,0280**

tableau (2) : la R-précision , la MAP et la précision moyenne de la formule I et de la thématique

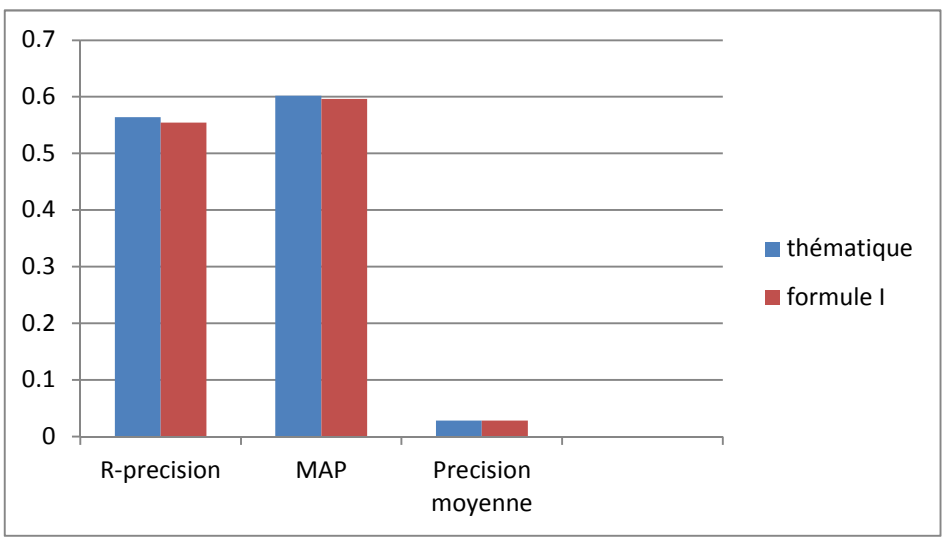


Figure (15) : - Comparaison de la MAP, R-précision et Précision moyenne du score thématique et de celui de la formule I-

Nous remarquons que notre approche n'apporte pas d'amélioration en ce qui concerne la **R-précision,** elle diminue de 0,0092,et aussi **la MAP** diminue de 0.0055 et pour la **précision moyenne** celle-ci a dégradé de 0,0002.

- **Le rappel, la précision et la f-mesure**

$$Précision = \frac{le\ nombre\ de\ documents\ pertinents\ retournés}{le\ nombre\ de\ documents\ retournés}$$

$$Précision = \frac{\mathbf{540}}{\mathbf{1656}}$$

$$Précision = \mathbf{0,3261}$$

$$Rappel = \frac{le\ nombre\ de\ documents\ pertinents\ retournés}{le\ nombre\ de\ documents\ pertinents}$$

$$Rappel = \frac{\mathbf{540}}{\mathbf{810}}$$

$$Rappel = \mathbf{0,6666}$$

$$F - mesure = \frac{2 \times rappel \times précision}{rappel + précision}$$

$$F - mesure = \frac{2 \times 0{,}6666 \times 0{,}3261}{0{,}6666 + 0{,}3261}$$

$$F - mesure = 0{,}4379$$

Où :

$$Précision = \mathbf{0,3261}$$

$$Rappel = \mathbf{0,6666}$$

$$F - mesure = \mathbf{0,4379}$$

- **Rappel interpolé :**

	Thématique	Formule I
iprec_at_recall_0.00	**0.6357**	**0.6282**
iprec_at_recall_0.10	**0.6351**	**0.6282**
iprec_at_recall_0.20	**0.6323**	**0.6259**
iprec_at_recall_0.30	**0.6296**	**0.6241**
iprec_at_recall_0.40	**0.6219**	**0.6169**
iprec_at_recall_0.50	**0.6205**	**0.6158**
iprec_at_recall_0.60	**0.5886**	**0.5824**
iprec_at_recall_0.70	**0.5846**	**0.5784**
iprec_at_recall_0.80	**0.5746**	**0.5683**
iprec_at_recall_0.90	**0.5717**	**0.5653**
iprec_at_recall_1.00	**0.5699**	**0.5636**

tableau (3) : comparaison du rappel interpolé de la thématique et celui de la formule I.

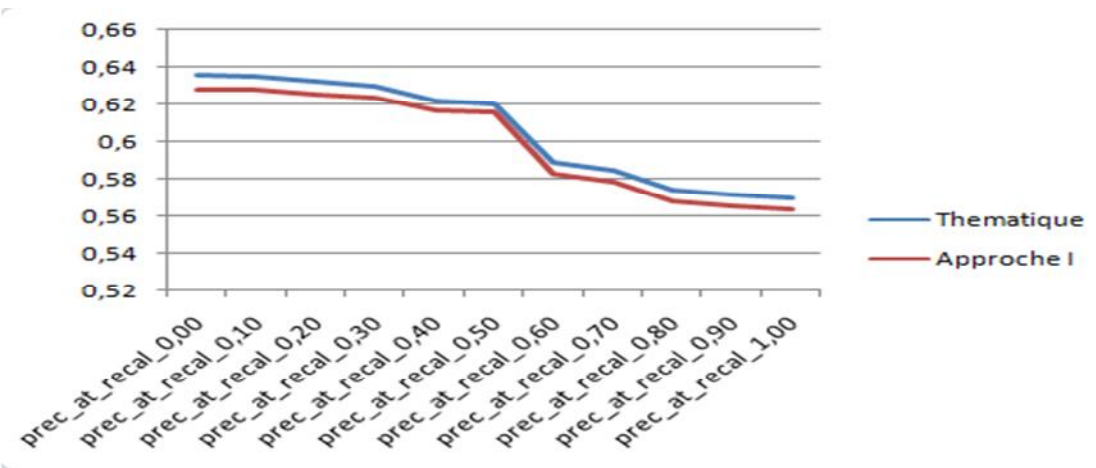


Figure (16) : - Courbe du rappel interpolé du score thématique et du score de la formule I.-

→Résultats

Après de diverses expérimentations nous constatons que la formule I reconnait une certaine satisfaction au niveau de la précision réelle mais n'apporte guère de résultats améliorés par rapport à la thématique au niveau autres mesures

Formule II

- **La précision@X**

	P@5	P@10	P@20
Thématique	0,1693	0,0919	0,0477
formule II	0,1693	0,0919	0,0477

tableau (4) : - P@5 , P@10, P@20 pour la thématique et la formule II -

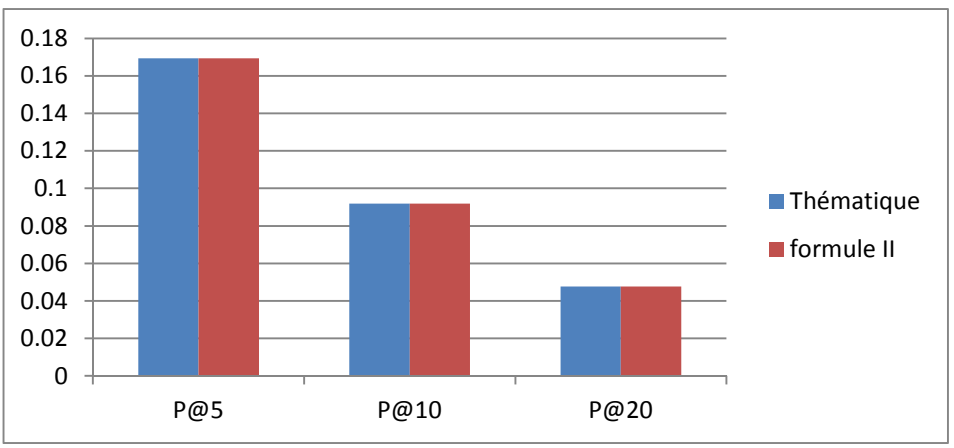


Figure (17) : - Comparaison des Précision@X du score de la formule II et de celui de la thématique -

Depuis ces résultats, nous en déduisons qu'avec la multitude de documents retournés pertinence se perd, et notre deuxième formule ne donne pas de résultats inférieurs à ceux de la thématique, par contre aussi elle ne les améliore pas.

- **La R-précision, la MAP et la précision moyenne**

	R-précision	MAP	Précision moyenne
La thématique	**0,5639**	**0,6018**	**0,0282**
L'approche II	**0,5251**	**0,5764**	**0,0270**

Tableau (5) : la R-précision, la MAP et la précision moyenne de la formule II et de la thématique.

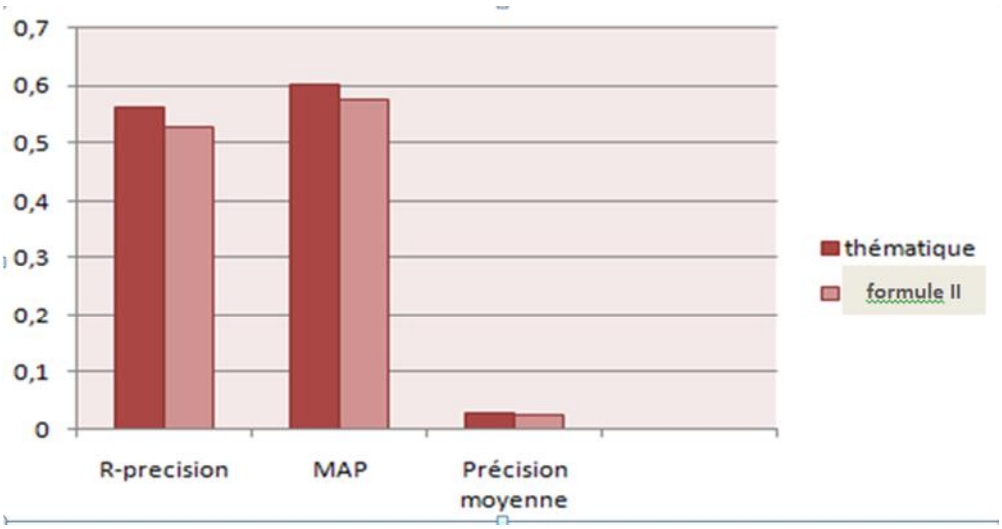


Figure (18) :- Comparaison de la MAP, R-précision et Précision moyenne du score thématique et de celui de la formule II

la **R-précision** diminue de 0,0388 ,**la MAP** diminue de 0.0254 et pour la **précision moyenne** celle-ci a dégradé de 0,012.

- **Le rappel, la précision et la f-mesure**

Précision= ***0.6666***

Rappel = ***0.3261***

F-mesure = ***0.4379***

- **Rappel interpolé :**

	thématique	**Approche II**
iprec_at_recall_0.00	0.6357	0.6092
iprec_at_recall_0.10	0.6351	0.6086
iprec_at_recall_0.20	0.6323	0.6062
iprec_at_recall_0.30	0.6296	0.6035
iprec_at_recall_0.40	0.6219	0.5951
iprec_at_recall_0.50	0.6205	0.5940
iprec_at_recall_0.60	0.5886	0.5633
iprec_at_recall_0.70	0.5846	0.5592

iprec_at_recall_0.80	0.5746	0.5493
iprec_at_recall_0.90	0.5717	0.5463
iprec_at_recall_1.00	0.5699	0.5446

Tableau(6) : comparaison du rappel interpolé de la thématique et celui de la formule II.

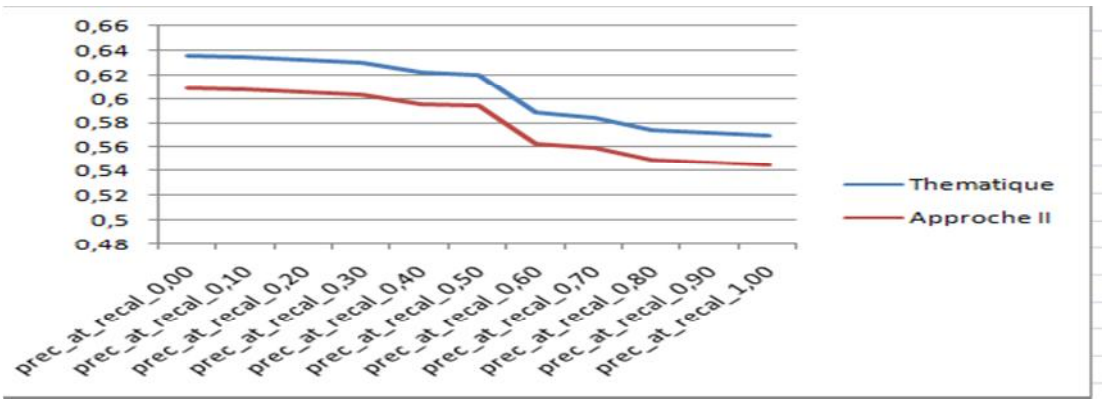


Figure (19) : Courbe de rappel interpolé de la thématique et celui de la formule II.

➔ **Résultats :**

En conséquent, nous constatons qu'après de nombreuses expérimentations faites sur la collection d'Edgard, la formule II aussi à son tour n'apporte pas de résultats attendus par rapport à ceux de la thématique. Et cela est dû à la taille de la collection que nous avons utilisée.

Conclusion

A travers ce chapitre nous avons présenté l'évaluation expérimentale de nos deux approches proposées, cette dernière est menée sur une petite collection de tweets appelée Tweets_Edgar, qui a altéré les résultats de nos deux approches, qui n'ont pas abouti aux résultats attendus.

Conclusion générale

Conclusion générale

Notre travail s'inscrit dans le cadre de la recherche d'information sociale, dans le contexte de microblogging pour la plateforme Twitter. Pour cela, nous avons commencé par présenter la recherche d'information classique, et expliqué les différentes phases de recherche et traité des modèles de la RI.

Nous avons axé notre travail d'un état de l'art selon deux dimensions différentes: la temporalité ainsi que les signaux sociaux, en présentant un aperçu des différentes approches les exploitant.

Dans le but de répondre à la problématique à laquelle nous nous sommes intéressées depuis le début de ce mémoire, qui réside à l'intégration et l'exploitation des informations sociales afin d'améliorer le processus de la recherche autrement dit, de régler le problème de difficulté de trouver une information pertinente, notamment avec le web actuel et l'émergence des réseaux sociaux. Nous avons proposé une solution incluant le facteur temporel et les commentaires. Elle calcule un score de pertinence en combinant linéairement le score thématique et le score social qui exploite comme informations sociales, le nombre de commentaires. Cette approche à un certain moment elle a amélioré la précision@X pour les 10 premiers documents retournés par rapport à la thématique. Mais nous avons remarqué une dégradation au niveau des autres mesures d'évaluation. En outre, nous y avons proposé une solution qui n'est qu'une reformulation de l'approche afin d'éviter que le SRI retourne les documents non pertinents, c'est-à-dire quand la thématique est nulle, le document est non pertinent ce qui implique que tout le score sera annulé et aucun document ne sera retourné par le SRI.

Après une multitude de tests expérimentaux, les résultats ne sont pas satisfaisants, c'est peut-être à cause de la taille petite de la collection utilisée ; elle n'a pas contribué à avoir une bonne amélioration avec nos deux propositions.

Limites et perspectives

Ce travail nous a permis d'approfondir nos connaissances déjà acquises dans le domaine de la RI soit du côté pratique ou bien théorique.

Nous avons réussi à proposer une solution dans le cadre de la recherche d'information dans Twitter, en intégrant la temporalité ainsi que les signaux sociaux.

Nous avons réussis l'implémentation et l'évaluation selon de diverses métriques de notre proposition, mais celle-ci n'a pas réussi à aboutir à des résultats satisfaisants.

Comme perspectives :

- Nous proposons de tester notre approche sur une collection de grande taille telle que TREC
- Combiner tous les signaux sociaux, autres que les commentaires par exemple les vues, les j'aimes…etc, et cela en gardant le même principe.
- Ou de combiner entre notre approche et celle de Mr DJEDDI (autrement dit entre les commentaires et les retweets)

Bibliographie

Bibliographie

[1] Jian-Yun Nie (Université de Montréal)

[2] R. A. Baeza-Yates, B A. Rebeiro-Neto. Modern Information Retrieval. ACM Press. Addison-Wesley, 1999.

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

[4] G. Salton, Automatic Information Organization and Retrieval. New York,McGraw. Hill Book Comapany, 1968

[5] HAMMACHE. A: « Recherche d'Information: un modèle de langue combinant mots simple et mots composés ». Thèse doctorat, UMMTO, 2013.

[6] Roberston, the probability Ranking principle in IR. Journal of documentation, 1977.

[7] Mandl, T. (2007). Recent developments in the evaluation of information retrieval systems: Moving towards diversity and practical relevance.

[8] Firas Damak. Etude des facteurs de pertinence dans la recherche de microblogs. PhD thesis, 2014

[9] Weng, J., Lim, E.-P., Jiang, J., et He, Q. (2010). Twitterrank : finding topic sensitive influential twitterers. In Wsdm'10 : Proceedings of the third acm international conference on web search and data mining (pp. 261–270). New York, NY, USA : ACM.

[10] Vosecky, J., Leung, K. W.-T., et Ng, W. (2012). Searching for quality microblog posts : Filtering and ranking based on content analysis and implicit links. , 397-413.

[11] Metzler, D., et Cai, C. (2011). USC/ISI at TREC 2011 : Microblog Track (Notebook Version). In *TREC'11: 20th Text Retrieval Conference*. National Institute of Standards and Technology (NIST).

[12] Duan, Y., Jiang, L., Qin, T., Zhou, M., et Shum, H.-Y. (2010). An empirical study on learning to rank of tweets. In Proceedings of the 23rd international conference on computational linguistics (pp. 295–303).

[13] Magnani, M., Montesi, D., et Rossi, L. (2012). Conversation retrieval for microblogging sites. Inf. Retr., 15 (3-4), 354-372.

[14] Zhao, L., Zeng, Y., et Zhong, N. (2011). A weighted multi-factor algorithm for microblog search. In Proceedings of the 7th international conference on active media technology (pp. 153–161). Berlin, Heidelberg : Springer-Verlag.

[15] Ounis, I., Lin, J., et Soboroff, I. (2011). Overview of the TREC-2011 Microblog Track. In TREC'11 : 20th Text Retrieval Conference.

[16] Robertson, S. (2004). « Understanding inverse document frequency : On theoretical arguments for idf. *Journal of Documentation*, *60* , 2004. »

[17] SEKOUR.M « Exploitation des signaux sociaux de Twitter pour améliorer la recherche d'information », mémoire Master, UMMTO, 2019

[18] HANNACHI.F LADAOUI.S «Recherche d'information dans Twitter proposition d'uneapproche de recherche d'influenceurs», mémoire Master, UMMTO, 2019

[19] DJEDDI.A et BENDOU.A «Recherche d'information temporelle dans les microblogs» mémoire Master, UMMTO, 2016.

[20] Lamjed BEN JABEUR and Lynda Tamine et Mohand Boughanem. Un modèle de recherche d'information sociale dans les microblogs : cas de twitter. *Marami*, 2012.

[21] Savonnet M. Frame A. Azaza L. Kirgizov S. «Evaluation de l'influence sur Twitter : Application au projet «Twitter aux Elections Européennes 2014 ». Mai 2015.

[21] MASAKI AONO, Abu Nowshed Chy, and Md Zia Ullah. A time and context aware reranker for microblog retrieval. The 29th Annual Conference of the Japanese Society for Artificial Intelligence, 2015.

Webographie

[A] http://www.iro.umontreal.ca/~nie/IFT6255/Introduction.pdf

[B] https://hal.archives-ouvertes.fr/tel-01110721/document

[C] https://fr.wikipedia.org/wiki/Twitter