République Algérienne Démocratique et Populaire
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique
UNIVERSITE MOULOUD MAMMERI DE TIZI-OUZOU

Faculté de Génie Electrique et Informatique
Département d'Informatique

Mémoire de Fin d'Etudes de MASTER ACADEMIQUE

Domaine : Mathématiques et Informatique

Filière : Informatique

Spécialité : Conduite de Projets Informatiques

Thème :

Recherche d'information dans Twitter : Proposition d'une approche de recherche d'influenceurs

Présenté par :

SAIDANI Djouher
HADJ RABAH Malika

Devant le Jury composé de:

Président : Mr S. SADI
Examinateur : Mme R. AOUDJIT

Encadré Par : Mme F.AMIROUCHE

PROMOTION : 2019-2020

Table des matières

Table des matières

TABLE DES MATIERES

TABLE DES MATIERES

Liste de figures

Liste de tableaux

Résumé :

Notre travail s'inscrit dans le domaine de la recherche d'information (RI) dans les microblogs, plus particulièrement dans la recherche d'influence dans la plateforme Twitter. En effet, étant donné Twitter une plateforme de prédilection, elle est convoitée par plusieurs bloggeurs où chacun essaye d'impacter et d'influencer sur la société par le biais de ses posts. Nous avons proposé une approche pour la mesure de l'influence basée sur la relation de retweet pour définir nos deux ratios de retweet(*tweets retweetés/tweets publiés*) et de somme de retweet (*somme de retweets/tweets publiés*), et nous avons exploité le h-index qui est un indicateur bibliométrique utilisé pour donner l'impact d'un chercheur. Par la suite nous avons combiné les deux ratios avec le h-index en un seul score, dit score d'influence. L'évaluation de notre approche repose sur un modèle de recherche qui intègre la mesure d'influence dans le calcul de pertinence des tweets.

Mots clés : recherche d'information, influenceur, influence, Twitter, microblogs.

Abstract :

Our work is a part of the field of social information retrieval (SIR), more particularly in the search for influencers in the Twitter platform. Indeed, given Twitter a very popular platform, it is coveted by several bloggers where everyone tries to impact and influence the society through his posts. In this context we have contributed to this problem by proposing an approach for the measurement of the influence which is based on the retweet relationship to define our ratios of retweet(tweets retweeted / tweets published) and sum of retweet (sum of retweets / tweets published), and we exploited the h_index which is a bibliometric indicator used to give the impact of a researcher. We then combined the two ratios with the h_index into a single score, called the influence score. The evaluation of our approach is based on a retrieval model that integrates the measure of influence into the calculation of tweets relevance.

Keywords : information retrieval, influencer, influence, Twitter, microblogs, social information retrieval.

Remerciements

Nous remercions en premier lieu, « **ALLAH** » le tout-puissant et le miséricordieux qui nous a donné la force, le courage et la patience pour réaliser ce travail.

En second lieu, nous remercions nos familles pour leur soutien moral durant toute la période du travail.

Nous tenons à remercier notre promotrice M^{me} AMIROUCHE F. pour son aide, ses conseils et sa confiance au cours de l'élaboration de ce travail.

Nous tenons également à exprimer notre gratitude et nos profonds respects envers tous les enseignants du département d'informatique de l'université de Mouloud MAMERI de Tizi-Ouzou (U.M.M.T.O.).

Nous remercions aussi, les membres du jury qui ont accepté d'évaluer ce travail et qui nous feront le plaisir de détecter nos erreurs et nous corriger.

Enfin, nous remercions toutes les personnes qui ont contribué de près ou de loin à l'élaboration de ce travail.

Dédicaces

A nos chers parents,

A nos frères et sœurs,

A nos familles,

A nos amis (es).

Introduction générale

Introduction générale

Le réseau Internet réunit aujourd'hui environ 4,5 milliards d'internautes, soit près de 55% de la population mondiale, selon des statistiques récentes de l'union International des télécommunications (ITU).

Parler d'internet aujourd'hui, c'est forcément parler des réseaux sociaux. Ces derniers, sont des services web délimités par un système, permettant aux individus de construire un profil public ou semi-public et d'articuler avec d'autres utilisateurs créant un réseau pour partager des informations, photos, vidéos ou carrément des relations et des centres d'intérêt. Ces véritables médias sociaux se distinguent par leur utilité (tantôt personnel, tantôt professionnel ou rencontres et autres...), leur logo et leurs audiences. Parmi ces différentes plateformes, nous citons à titre d'exemple Facebook, Twitter, Instagram et LinkedIn...

Notre travail s'intéressera de près au réseau Twitter. Ce dernier, est un réseau social de microblogage géré par l'entreprise Twitter Inc. Il permet à un utilisateur d'envoyer gratuitement de brefs messages, appelés tweets, sur internet, par messagerie instantanée ou par SMS. Twitter compte plus de 300 millions d'utilisateurs actifs par mois et plus de 500 millions de tweets envoyés par jour.

Il est remarquable que les usagers de Twitter produisent beaucoup de contenu ; des informations fraîches, pertinentes et moins importantes dans certaines situations. Mettre la main ou cerner ce contenu s'avère une tâche difficile car la qualité d'une information est associée aux messages publiés d'un côté, ainsi qu'à l'importance de son auteur (influenceur), nous parlons ici de la notion d'influence qui représente l'impact d'un utilisateur sur les autres. Notre travail reposera justement sur la recherche d'information (RI) dans Twitter et propose une approche de recherche d'influenceurs.

Contribution :

Dans notre travail, nous proposons une approche qui permet de mesurer l'influence dans Twitter. Nous nous sommes basé sur la relation de retweet pour définir nos deux ratios de retweet (*tweets retweetés/ tweets publiés*) et de somme de retweet (*somme de retweets/tweets publiés*), et nous avons exploité le h-index, ce dernier est un indicateur bibliométrique utilisé pour donner l'impact d'un chercheur. Par la suite, nous avons combiné les deux ratios avec le h-index en un seul score, dit score d'influence.

Organisation du mémoire :

Le présent travail s'articule autour de quatre (04) chapitres :

Chapitre 1 : Généralités sur la recherche d'information, dans ce chapitre nous présentons les différents concepts liés à la recherche d'information, puis nous présentons la plateforme de microblogging Twitter et nous introduisons les spécifiés de la recherche d'information dans twitter.

Chapitre 2 : Les mesures de l'influence sociale dans Twitter. Ce chapitre est un état de l'art sur les différentes approches déjà proposées dans le cadre de la mesure de l'influence.

Chapitre 3 : L'approche proposée. Dans ce chapitre, nous expliquons l'approche que nous avons proposée.

Chapitre 4 : Implémentation et évaluation, ce dernier chapitre présente les détails d'implémentation et de mise en œuvre de notre approche, ainsi que les résultats de son évaluation.

Nous terminons notre mémoire par une conclusion générale.

Chapitre 01 :

Généralités sur la Recherche d’information

Introduction :

La recherche d'information (RI) est le processus par lequel les informations (ou les documents qui les contiennent) sont stockées, puis mises à la disposition des utilisateurs à la demande. Ce chapitre a pour objectif de présenter la recherche d'information (RI). Il est organisé en deux parties : Dans la première partie, nous présentons les concepts de base de la RI en général. Dans la seconde partie, nous nous intéresserons à la RI dans Twitter en particulier.

1. Recherche d'information :

La recherche d'information est une discipline relativement ancienne qui a pour but de retrouver, à partir d'une collection de documents existants, les documents pertinents qui répondent au mieux au besoin informationnel de l'utilisateur formellement exprimé par une requête. La recherche d'information (RI) traite de la représentation, du stockage, de l'organisation et de l'accès à l'information [Manning et al., 08]. La RI intègre des modèles, des méthodes, des procédures et des techniques dont le but est de faciliter l'accès à l'information pertinente pour un utilisateur ayant un besoin en information [Daoud, 09].

La RI est mise en œuvre à travers des **Systèmes de Recherche d'Information** (SRI). Ces systèmes ont pour but de mettre en œuvre des techniques et des moyens permettant de retourner les documents pertinents d'une collection en réponse à un besoin en information d'un utilisateur, exprimée par une requête [Baeza-Yates. 11].

2. Notions de base d'un système de recherche d'information :

2.1 Besoin en information et Requête :

Le besoin en information est une expression mentale des informations que l'utilisateur recherche. Cette notion est souvent assimilée au besoin de l'utilisateur.

La requête est une expression approximative du besoin en information de l'utilisateur, elle représente l'interface entre le SRI et l'utilisateur. Elle est exprimée par un ensemble de mots clés, mais elle peut être exprimée en langage naturel, booléen ou graphique.

2.2 Document et Collection de documents :

Un document est un élément essentiel dans un SRI. C'est un support physique de l'information, qui peut être du texte, une page web, une image, une séquence vidéo, etc.

Une Collection de documents (Fond documentaire ou corpus) constitue l'ensemble des informations accessibles et exploitables par le SRI.

2.3 Pertinence :

La pertinence est une notion importante dans la RI, et est l'objet de tout système de recherche d'information. Elle définit la correspondance entre un document et une requête utilisateur, ou encore le degré de « relation » du document à la requête. Essentiellement, deux types de pertinences sont définies : La pertinence système et la pertinence utilisateur.

— Pertinence système **:**

Le système mesure un degré de pertinence, ou valeur de similarité (ou score), entre un document et une requête afin d'évaluer l'adéquation du contenu des documents vis-à-vis de celui de la requête. Ce type de pertinence est objectif et déterministe.

— Pertinence utilisateur **:**

Elle se traduit par les jugements de pertinence de l'utilisateur sur les documents fournis par le SRI en réponse à une requête. La pertinence utilisateur est subjective, car deux utilisateurs qui ont des centres d'intérêts différents peuvent juger différemment un même document envoyé pour une même requête. De plus, cette pertinence est évolutive : un document jugé non pertinent à l'instant t pour une requête donnée, peut être jugé pertinent à l'instant t+1, car la connaissance de l'utilisateur sur le sujet a évolué.

3. Processus de la RI :

Le système de recherche d'information (SRI) intègre trois fonctions principales : l'Indexation, l'appariement document- requête et la reformulation de requête, qui sont mises en œuvre dans le processus en U de la recherche, comme illustré dans la Figure 1.1.

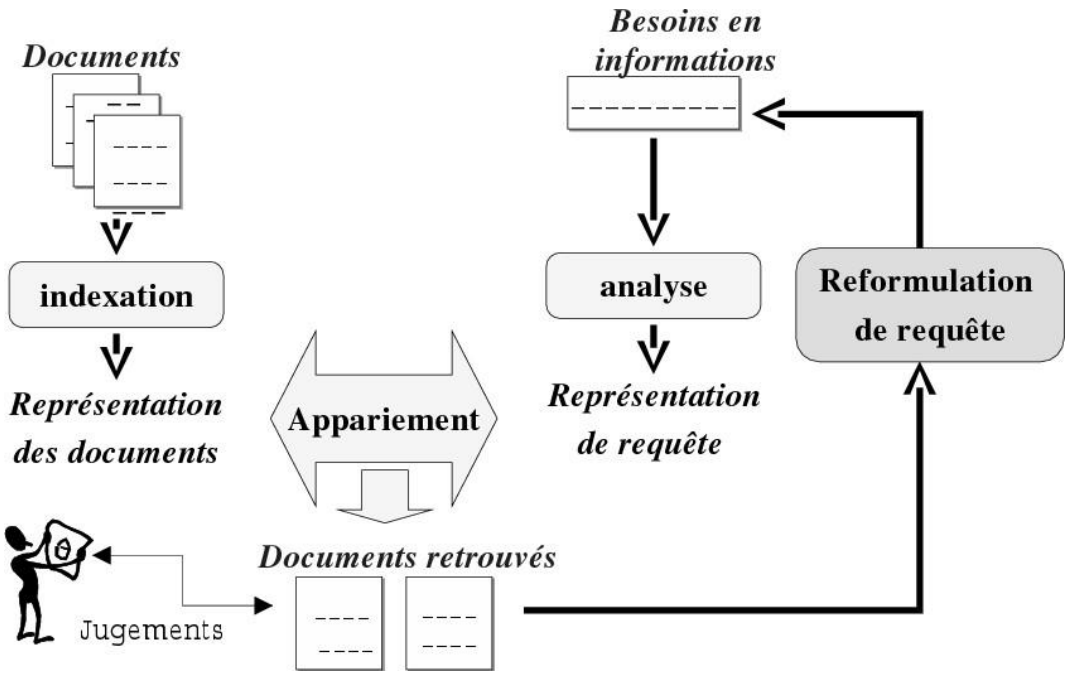


Figure 1.1 : Processus en U de la Recherche d'Information

3.1 Indexation :

L'indexation est une opération fondamentale dans le processus de RI. Elle consiste à analyser les documents et la requête utilisateur, dans le but de construire leurs représentations internes, ou index, en vue de faciliter la recherche. L'index est censé représenter au mieux le contenu sémantique du document.

L'indexation peut-être manuelle, semi-automatique ou automatique [Salton, 88 ; Salton et al., 88].

3.1.1 L'indexation Manuelle :

Chaque document de la collection est analysé par un expert du domaine ou par un documentaliste qui se charge de définir les mots-clés les plus significatifs pour représenter le document. Ce mode d'indexation est fondé sur le jugement humain.

3.1.2 L'indexation Semi-automatique :

L'indexation semi-automatique combine les deux types d'indexation manuelle et automatique. La tâche d'indexation est réalisée conjointement par un programme informatique et un spécialiste du domaine. Dans ce cas, chaque document est d'abord analysé à l'aide d'un

processus entièrement automatisé. Le choix final des descripteurs revient à l'indexeur humain [Hammache, 13].

3.1.3 L'indexation Automatique :

L'indexation automatique [Boubekeur, 08] consiste à analyser chaque document à l'aide d'un processus entièrement automatisé. Elle est réalisée par un programme informatique et s'appuie sur un ensemble d'étapes (figure 1.2) comme suit :

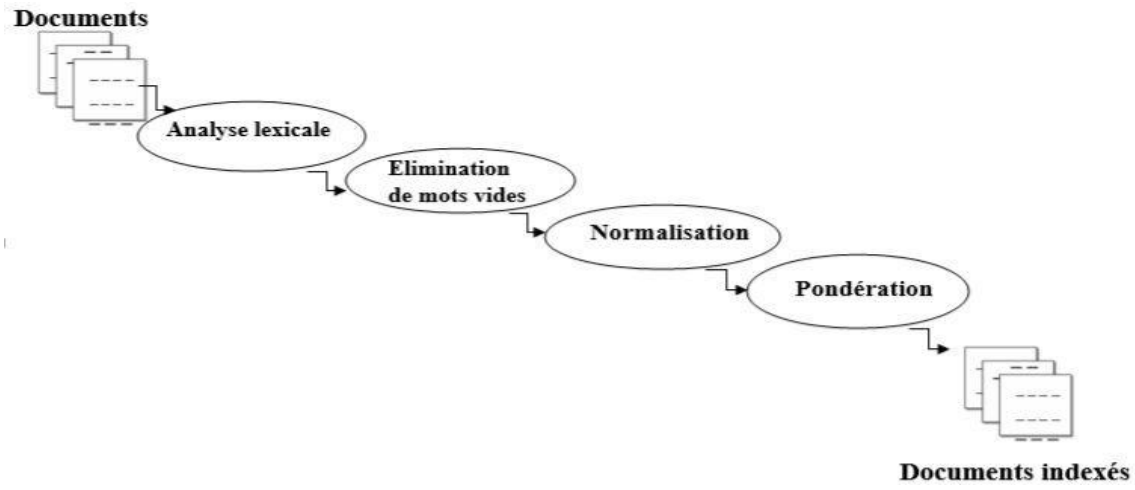


Figure 1.2 : Étapes de traitements pendant l'indexation automatique

— Analyse lexicale :

Ce processus consiste à découper un texte de document, en un ensemble de termes ou unités lexicales. Le découpage se base sur des caractères de séparation (blancs, ponctuation, …).

— Élimination de mots vides :

Cette étape consiste à éliminer les mots non significatifs (pronom personnel, prépositions,…) dans le document. On distingue deux techniques pour l'élimination des mots vides [Hammache, 13] :

- Utilisation d'une liste préétablie de mots vides (anti-dictionnaire ou Stop-List).
- L'élimination des mots ayant une fréquence qui dépasse un certain seuil dans la collection.

L'élimination de mots vides permet de réduire l'index (gain d'espace mémoire) et donc améliorer le temps de réponse du système.

— Normalisation :

Ce processus consiste à retrouver pour un mot sa forme normalisée [K.Mechach, 16]. L'idée qui conduit à utiliser la normalisation, est de pouvoir indexer un ensemble de mots de la même famille, par un seul mot (format unique) dit lemme ou racine qui représente le même concept. Plusieurs stratégies de normalisation sont utilisées [Hammache. 13], on distingue :

- La lemmatisation (*stemming* en anglais) permettant de regrouper les différentes formes des mots d'une même famille en les réduisant à leur forme canonique ou lemme. Les verbes par exemple ont une forme canonique reconnaissable à leur infinitif, les noms, adjectifs, articles,… ont pour forme canonique le masculin singulier.
- La troncature appelée souvent 'désuffixation', consiste à éliminer les suffixes des mots significatifs du texte indexé et [Manning et al., 08] permettant de transformer des flexions en leur racine (ou radical).

— Pondération :

La pondération des termes est l'une des fonctions fondamentales en RI. Elle permet d'assigner une valeur numérique de représentativité (dite poids) à chaque terme t d'un document d ou d'une requête q. Le poids représente l'importance du terme dans le document. On distingue deux types de pondérations :

- Pondération locale basée sur la fonction Tf (Term Frequency) **:** mesure le nombre d'occurrences (la fréquence) du terme t_i dans le document d_j. Un bon terme est un terme qui apparait fréquemment dans le document. Les fonctions de pondération locale les plus utilisées sont :

$$\mathrm{tf} = \begin{cases} freq(t,d) \\ 1 + \log(freq(t,d)) \\ 0.5 + 0.5\left(\dfrac{freq(t,d)}{\max(freq(t,d))}\right) \end{cases}$$

Où :

- tf : est la fréquence d'occurrences du terme t_i dans le document d_j ;
- freq(t,d) : Est le nombre d'occurrences de chaque terme t dans chaque document d.
- Pondération Globale basée sur la fonction idf (Inverse Document Frequency) **:** mesure l'importance d'un terme dans toute la collection. Un poids plus important est donné au terme qui apparait dans peu de documents. Ce poids est inversement proportionnel à la fréquence documentaire du terme dans la collection. Il est alors mesuré par *Idf (inverted df) ou* fréquence documentaire inverse. Il est exprimé par l'une des formules suivantes :

$$idf = \begin{cases} \log\left(\frac{N}{ni}\right) \\ \log\left(\frac{N - ni}{ni}\right) \end{cases}$$

Avec :

- N : Le nombre total de documents de la collection ;
- ni : La fréquence en document du terme ti.
- Pondération TF-IDF : Le calcul des poids des termes se base aussi bien sur la pondération locale et globale. Dans ce cas, le terme choisi combine deux caractéristiques : il est important dans le document et il apparaît peu dans les autres documents. C'est cette mesure qui est utilisée le plus souvent en RI. Elle s'exprime comme suit : $tf * idf$

Un autre facteur est utilisé par la majorité des modèles pour normaliser la pondération : c'est la taille du document. Plus ce dernier est long, plus le terme est fréquent, ce qui a pour effet de favoriser les documents longs par rapport aux documents plus courts dans une recherche. Pour éviter ce problème, les poids sont normalisés en tenant compte de la longueur des documents.

3.2 L'appariement requête-document :

Le processus d'appariement permet de retrouver les documents pertinents qui répondent à la requête utilisateur, après avoir calculé un score de correspondance (ou degré de pertinence) entre la représentation de chaque document et celle de la requête. Ce score est donné par une fonction de similarité appelée *RSV (d, q)* (*Retrieval Status Value*), d est un

document et q est la requête. L'expression de la fonction d'appariement est tributaire du modèle de RI choisi.

Il existe deux types d'appariement :

— Appariement exact : le résultat est une liste de documents respectant exactement la requête spécifiée avec critères clairs. Les documents retournés ne sont pas triés.

— Appariement approché : le résultat est une liste de documents censés être pertinents pour la requête. Les documents retournés sont triés selon par degrés de pertinence décroissants.

3.3 La reformulation de requête :

L'utilisateur de moteurs de recherche n'est pas forcément un professionnel de la documentation. Il ne sait pas choisir les bons termes qui expriment le mieux ses besoins d'information. La reformulation de requête permet de générer une nouvelle requête plus adéquate que celle initialement formulée par l'utilisateur, en ajoutant des termes significatifs et/ou ré-estimant les poids des termes de la requête initiale.

On distingue principalement deux approches de reformulation de requête :

— L'expansion automatique de requête :

Dans le cas de l'expansion automatique de requête, la requête de l'utilisateur est remaniée automatiquement en rajoutant de nouveaux termes (issus de ressources linguistiques existantes ou de ressources construites à partir des collections) à la requête initiale.

— Réinjection de pertinence (Relevance Feedback) :

Le processus de réinjection de pertinence comporte principalement trois étapes :

- L'échantillonnage : cette étape permet de construire un échantillon de documents à partir des éléments jugés pertinents par l'utilisateur.
- L'extraction des évidences : cette étape consiste à sélectionner les termes pertinents qui serviront à l'enrichissement de la requête initiale.
- La réécriture de la requête : consiste à construire une nouvelle requête en combinant la requête initiale avec les informations extraites dans l'étape

précédente. Plusieurs approches ont été développées la plus reconnues est celle de Rocchio [J. Rocchio, 71] adaptée au modèle vectoriel (que nous détaillerons en section suivante).

4. Modèles de recherche d'information :

D'une façon générale, l'appariement requête-document et le modèle d'indexation permettent de caractériser et d'identifier un modèle de recherche d'information. Ce dernier fournit une interprétation théorique de la notion pertinence.

De nombreux modèles ont été proposés en RI, nous les présentons dans ce qui suit.

4.1 Le modèle Booléen :

Ce modèle est le premier modèle utilisé dans le domaine de la recherche d'information. Il est basé sur la théorie des ensembles et l'algèbre de Boole [Salton et al, 83]. Il s'est imposé grâce à la simplicité et à la rapidité de sa mise en œuvre. Dans ce modèle, une requête q est représentée sous forme d'une expression logique composée de mots-clés reliés par des opérateurs logiques (NOT, AND et OR). Un document d est représenté par un ensemble de termes d'indexation (ou mots-clés).

L'appariement entre une requête et un document est un appariement exact. Il est basé sur la présence/absence des termes de la requête dans les documents. La similarité entre un document et une requête est alors une fonction à valeurs 0 ou 1 (résultat binaire), qui peut être définie de la manière suivante :

$$RSV(d, \text{ti}) = 1 \text{ si ti} \in \text{d; 0 sinon;}$$

$$RSV(d, \text{qi} \wedge \text{qj}) = 1 \text{ si RSV (d, qi)} = 1 \text{ AND RSV (d, qj)} = 1\text{; 0 sinon;}$$

$$RSV(d, \text{qi} \vee \text{qj}) = 1 \text{ si RSV (d, qi)} = 1 \text{ OR RSV (d, qj)} = 1\text{; 0 sinon;}$$

$$RSV(d, \neg\text{qi}) = 1 \text{ si RSV (d, qj)} = 0\text{; 0 sinon;}$$

Bien que ce modèle soit simple à mettre en œuvre, il présente des inconvénients :

- Difficulté de la formalisation des requêtes.
- le modèle booléen se base, pour la sélection des documents, sur la pertinence exacte. Il considère comme non pertinents les documents qui ne contiennent pas tous les

termes de la requête utilisateur. Un document est donc soit pertinent, soit non pertinent.

- Dans ce modèle, tous les termes ont le même poids lorsqu'ils sont dans le document, alors que certains sont plus représentatifs que d'autres. Par conséquent les documents ne sont pas ordonnés

4.2 Le modèle vectoriel :

Le modèle vectoriel appelé aussi modèle algébrique, est proposé par Salton [Salton, G. 70]. Ce modèle représente les documents et les requêtes sous forme de vecteurs dans l'espace vectoriel engendré par tous les termes de la collection de documents [Salton et McGill. 86].

Dans ce modèle, un document est représenté par un vecteur $di = (W_{i1}, W_{i2}, \ldots, Wi_m)$ pour $i=1,2,\ldots, n$. Où W_{ij} est le poids du terme i dans le document, n le nombre de document, et la requête représentée par un vecteur $q = (Wq_1, Wq_2, \ldots Wq_n)$ où W_{iq} est le poids (souvent 0 ou 1) du terme i dans la requête. La similarité entre les vecteurs di et q, est exprimée par l'une des mesures suivantes :

— Le produit Scalaire :

$$RSV(q, dj) = \sum_{i=1}^{n} W_{iq} * W_{ij}$$

— La mesure de Cosinus :

$$RSV(q, dj) = \frac{\sum_{i=1}^{n} W_{iq} * Wij}{\sqrt{\sum_{i=1}^{n} W_{iq}^{2} * \sum_{i=1}^{n} W_{ij}^{2}}}$$

— La mesure de Dice :

$$RSV(q, dj) = \frac{2 * \sum_{i=1}^{n} W_{iq} * W_{ij}}{\sum_{i=1}^{n} W_{iq}^{2} + \sum_{i=1}^{n} W_{ij}^{2}}$$

— La mesure de Jaccard :

$$RSV(q, dj) = \frac{\sum_{i=1}^{n} W_{iq} * W_{ij}}{\sum_{i=1}^{n} W_{iq}^{2} + \sum_{i=1}^{n} W_{ij}^{2} - \sum_{i=1}^{n} W_{iq} * W_{ij}}$$

La figure 1.3 illustre une représentation vectorielle dans un espace composé de deux termes, avec deux documents (d1, d2) et une requête (q).

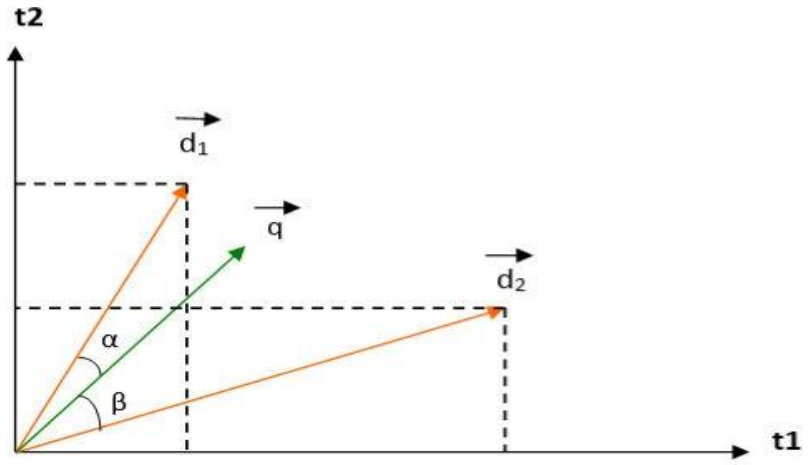


Figure 1.3 : Représentation vectorielle de deux documents et une requête

Le score de pertinence système de chaque document est proportionnel à la distance entre les représentations vectorielles du document et de la requête. Plus l'angle formé par le document et la requête est petit, plus le score de pertinence du document pour la requête est grand. Ce document est alors considéré pertinent à la requête. Dans le cas de la figure 1.3, le document d_1 est pertinent à la requête q.

4.3 Le modèle probabiliste :

Proposé par Maron et Kuhns au début des années 60 [Maron et Kuhns. 60], le modèle probabiliste est fondé sur le calcul de probabilité de pertinence d'un document pour une requête. La pertinence d'un document d et une requête q, notée RSV (d_i, q), est mesurée par le rapport entre la probabilité P $(per \mid q, d_i)$ que di soit pertinent (per) pour q, et la probabilité P $(Nper \mid q, d_i)$ qu'il soit non pertinent $(Nper)$:

$$RSV(d_i, q) = \frac{P(per \mid q, d_i)}{P(Nper \mid q, d_i)}$$

Ces probabilités sont estimées par les probabilités de présence ou d'absence d'un terme de la requête dans un document pertinent ou non pertinent. En appliquant la formule de Bayes pour ces deux probabilités, on obtient :

$$P(per\ /\ q, d_i) = \frac{P(per \mid q), P(d_i \mid per, q)}{P(d_i)}$$

$$P(per\ /\ q, d_i) = \frac{P(Nper \mid q), P(d_i \mid Nper, q)}{P(d_i)}$$

Où :

- $P(d_i)$: est la probabilité de choisir le document di, on considère qu'elle est constante ;
- $P(d_i \mid per, q)$: indique la probabilité que di fasse partie des documents pertinents pour la requête q ;
- $P(d_i \mid Nper, q)$: indique la probabilité que di, fasse partie des documents non pertinents pour la requête q ;
- $P(per \mid q)$ et $P(Nper \mid q)$: indique respectivement la probabilité de pertinence et de non pertinence d'un document quelconque.

4.4 Le modèle de langue :

Le modèle de langue (*Language Model*) ou modèle statistique de langue est un modèle probabiliste utilisé dans diverses applications de traitement automatique de langue : la reconnaissance de parole, la traduction automatique, la recherche d'information, etc. [Hammache. 13]. Contrairement aux modèles de recherche classique où l'estimation de degré de pertinence d'un document à l'égard d'une requête se résume à une opération d'appariement entre leurs termes, le modèle de langue estime que la pertinence d'un document pour une requête est en rapport avec la probabilité que la requête puisse être générée par le document.

Le principe de base de ce modèle est de construire un modèle de langue M_d pour chaque document d, puis de calculer la probabilité $P(Q \mid M_d)$ qu'une requête $Q = (t_1, t_2, t_3 \ldots t_n)$ puisse être générée par le modèle de langue du document. Cette probabilité de génération est exprimée comme suit :

$$RSV(d, Q) = P(Q|M_d) = P(Q = (t_1, t_2, t_3 \dots t_n)|M_d) = \prod_{i=1}^{n} P(t_i |M_d)$$

Où :

– n est le nombre de termes dans la requête et t_i est un terme de la requête, $(1 \leq i \leq n)$

Sachant que :

$$P(t_i|M_d) = \frac{tf(t_i, d)}{|d|}$$

Où :

– tf (t_i, d) est la fréquence du terme t_i dans le document d.

Afin de pallier le problème des termes de la requête absents des documents, (tf (ti, d)=0 conduit systématiquement à $P(Q|M_d)=0$), il s'avère essentiel d'utiliser des techniques de lissage (*Smoothing*). Le lissage consiste à assigner des probabilités non nulles aux termes qui n'apparaissent pas dans les documents.

5. Evaluation des systèmes de recherche d'information :

L'évaluation permet de paramétrer le modèle, d'estimer l'impact de chacune de ses caractéristiques et de fournier des éléments de comparaison entre les modèles.

5.1 Collections de test :

L'évaluation d'un SRI consiste principalement à mesurer ces performances sur la base d'une collection de test contrôlée et des métriques d'évaluation standards définis selon des critères d'efficacité. Chaque collection (ou corpus) de test est composée d'un ensemble de documents, d'un ensemble de requêtes et des jugements de pertinence associés à ces requêtes.

Il existe plusieurs compagnes d'évaluation, parmi elles, on trouve TREC : La compagne d'évaluation TREC (*Text REtrieval Conference*), co-organisée par NIST[1] et la DARPA, est créée en 1992. Elle a pour but d'encourager la recherche documentaire basée sur de grandes collections de test, tout en fournissant l'interface nécessaire pour l'évaluation des méthodologies de recherche d'information. Pour chaque session de TREC, un ensemble de documents et de requêtes est fourni. Les participants exploitent leurs propres systèmes de

[1] https : //trec.nist.gov/

recherche sur les données et renvoient au NIST une liste ordonnée de documents. Le NIST évalue ensuite les résultats. L'évaluation se base sur un ensemble de mesures d'évaluation.

5.2 Mesures d'évaluation :

L'objectif principal de tous les systèmes de recherche d'information est de retrouver les documents pertinents et rejeter les documents non pertinents. Cet objectif est évalué par les mesures de précision et de Rappel. La Figure 1.4 présente les éléments de définition du rappel et de la précision.

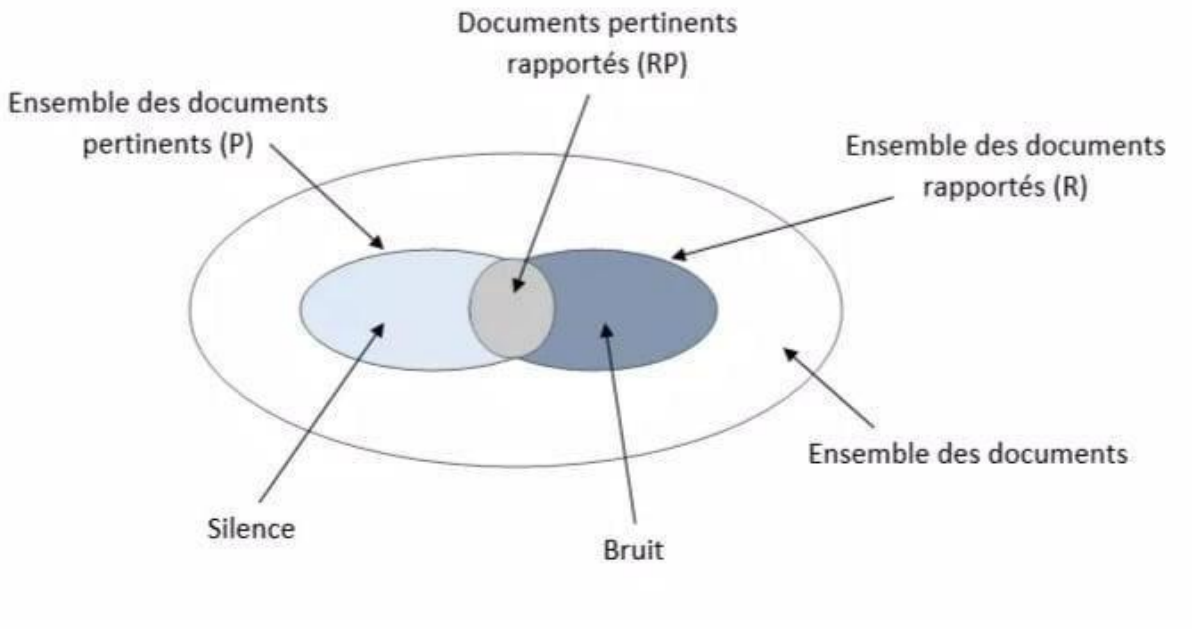


Figure 1.4 : Définition du Rappel et de la Précision

— Rappel :

Le rappel mesure la capacité du système à retrouver tous les documents pertinents répondants à une requête. Elle mesure la proportion des documents pertinents restitués par le système relativement à l'ensemble des documents pertinents contenus dans la base documentaire.

$$Rappel = \frac{Nombre\ de\ documents\ pertinents\ retournés}{Nombre\ de\ documents\ pertinents\ total} = \frac{RP}{P}$$

— Précision :

La précision mesure la capacité du système à rejeter tous les documents non pertinents à une requête donnée par le rapport entre l'ensemble des documents sélectionnés pertinents et l'ensemble des documents sélectionnés. La précision mesure la proportion de documents pertinents retrouvés parmi tous les documents retrouvés par le système.

$$Précision = \frac{\text{nombres de documents pertinets retournés}}{\text{nombre total de documents retournés par le système}} = \frac{RP}{R}$$

— Courbe rappel-précision :

La courbe illustrée dans la figure 1.5 montre qu'il est toujours possible d'obtenir une précision élevée au prix d'un rappel faible, ou un rappel élevé au prix d'une précision faible.

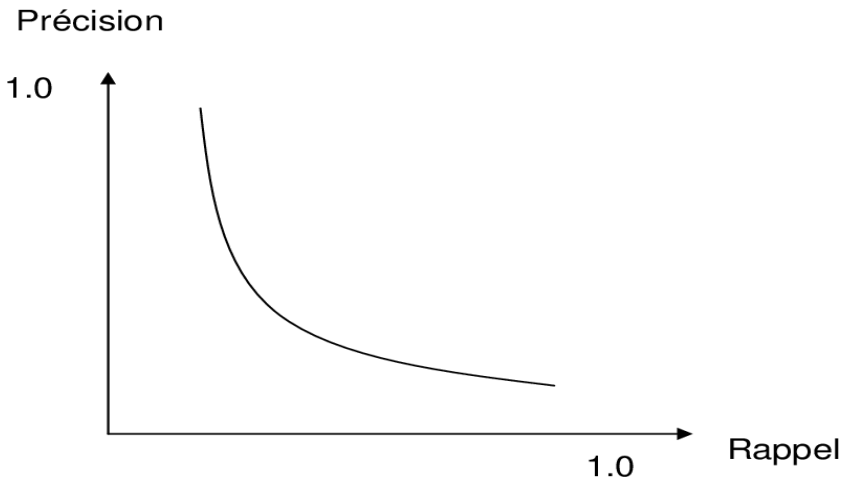


Figure 1.5 : Courbe rappel-précision

D'autres mesures complémentaires au rappel et à la précision peuvent être utilisées, dont :

— Bruit :

Ensemble de documents non pertinents retournés par les SRI en réponse à une requête donnée.

$$Bruit = 1 - P$$

— **Silence :**

Ensemble de documents pertinents non sélectionnés par le système lors d'une recherche d'information.

$$Silence = 1 - R$$

— R-précision :

Cette précision mesure la proportion des documents pertinents retrouvés après que R documents ont été retrouvé, Où R est le nombre de documents pertinents pour la requête considérée.

— MAP (*Mean Average Precision*) :

C'est la moyenne des précisions moyennes (*Average precision-AP*) obtenues sur l'ensemble des requêtes à chaque fois qu'un document pertinent est retrouvé.

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

Avec :

- Q : est l'ensemble des requêtes ; |Q| : est le nombre total de requêtes ;
- AP_q : est la moyenne des précisions à chaque rang de document pertinent pour la requête q.

6. La recherche d'information dans les microblogs (cas Twitter) :

La recherche d'informations sociales (*social information retrieval*) est un domaine d'étude relativement nouveau autour de la recherche et de l'acquisition d'informations à partir des espaces sociaux sur Internet. On s'intéresse dans ce qui suit à la recherche d'information dans le réseau social Twitter.

6.1 Présentation de Twitter :

Twitter est le site de microblogging le plus connu dans le monde. Twitter a été créé en mars 2006 par Jack Dorsey à San Francisco. Il a été lancé par la suite au mois de juillet de la

même année. Depuis lors, Twitter a évolué exponentiellement [Damak. 14]. Il est aujourd'hui un réseau social incontournable du paysage média. 500 millions de tweets de 280 caractères y sont échangés tous les jours. Ainsi, au troisième trimestre 2019, Twitter comptait 145 millions d'utilisateurs actifs quotidiens et 330 millions d'utilisateurs actifs mensuels.

Figure 1.6 : Logo de twitter.

6.1.1 Vocabulaire de Twitter :

— Twitto : c'est l'utilisateur de twitter (ou bloggeur).

— Tweet (gazouillis): un message publié via Twitter, qui peut contenir des textes libres, des images ou vidéos, des liens, des hashtags et/ou des mentions, ne dépassant pas 280 caractères.

— Timeline : liste des flux personnels.

— Follower (abonné) : lorsqu'un twitto s'abonne à un profil sur Twitter, il en devient un follower, il reçoit ainsi les tweets de cette personne dans son fil d'actualités. Tout twitto peut ainsi être un follower et posséder des followers.

— Following (abonnements) : à l'inverse, cela désigne le nombre de personnes suivies par un twitto.

— Retweet : une des fonctionnalités les plus utilisées de Twitter. Elle permet de transférer le tweet d'un contact à d'autres pour augmenter sa visibilité.

— Hashtags : désigne un mot-clé précédé de #, il renvoie à un thème ou/et à des tweets.

— Mentions : consiste à identifier un autre compte dans un tweet, grâce au signe @.

— Replies : réponses à une personne en particulier en précisant son nom précédé du symbole « @ ».

6.1.2 Les interactions dans twitter :

Il existe quatre types de relations publiques sur Twitter : user-to-user, user-to-tweet, tweet-to-tweet et tweet-to-user. Le tableau 1.1 indique les actions valides pour chaque type de relation :

Tableau 1.1 : Relations entre les utilisateurs et les tweets sur Twitter

	User	Tweet
User	Follows (suit)/ Is followed by. Mention Replies to Retweets to	Posts Retweets Likes Replies
Tweet	Posted by Retweeted by Liked by Replied by	Replies/ Is replied from Retweets/ Is retweeted from

6.2 La recherche d'information dans twitter :

Twitter est l'exemple le plus populaire des plateformes de microblogging. Ces plateformes sont les réseaux sociaux les plus récents du Web 2.0. Elles sont considérées comme une nouvelle forme de blogs, où les informations diffusées sont courtes et publiées plus rapidement. Ces informations concernent différents sujets. Le nombre de requêtes soumises à Twitter correspond à 42 % des requêtes soumises à Google. Ce chiffre montre l'importance de Twitter en tant que source d'informations et la dépendance des utilisateurs à cette source d'information [Damak, 14]. Dans cette section, nous nous intéressons à la façon dont on accède à l'information sur Twitter, en particulier en tenant compte des facteurs de pertinence liés à ce réseau social, puis nous définirons l'évaluation de la RI dans Twitter.

6.2.1 Accès à l'information dans les microblogs :

— Recherche en temps réel : pour cette tâche, l'utilisateur souhaite obtenir de l'information pertinente la plus fraîche possible vis-à-vis d'un besoin en information, d'où la date de publication d'un document est considérée comme un facteur de pertinence très important dans cette recherche.

— Recherche de microblogueurs : c'est la recherche des utilisateurs les plus populaires, ils peuvent être des leaders, journalistes, influenceurs ou encore des débatteurs, ou des experts dans des domaines spécifiques.

— Détection d'opinions : l'objectif est de retrouver les documents exprimant des opinions sur le sujet de la requête. Par exemple, [Shamma et al., 09] ont trouvé que les tweets peuvent être utilisés pour annoter les débats politiques avec les opinions des téléspectateurs. Plus précisément, ils ont constaté que le taux de messages contenant des opinions dans Twitter peut servir comme un prédicateur de l'évolution des sujets dans l'événement médiatisé.

— Classification thématiques des microblogs **:** cette classification permet de classer les utilisateurs en fonction de leurs centres d'intérêts. Les sujets discutés dans les microblogs sont identifiés afin de créer des filtres thématiques sur les flux d'information. La première solution utilisée est de regrouper les microblogs en fonction des hashtags qu'ils contiennent.

— Détection des tendances : les tendances sont généralement des événements émergeants, les dernières nouvelles, les plus récentes. Elles indiquent des sujets qui captent l'attention du public. La détection des tendances revêt donc une grande utilité pour les journalistes et les analystes, car elle leur permet d'être rapidement actifs sur les sujets « tendances » [Damak, 14].

6.2.2 Facteurs de pertinence dans les microblogs :

a. Facteurs basés sur le contenu des tweets :

— Popularité du thème du tweet : si on trouve plusieurs tweets ayant le même contenu qu'un autre tweet, alors ce dernier est considéré comme étant populaire.

— Longueur du tweet : intuitivement, plus un message est long, plus il contient de l'information.

— Hashtags dans le tweet : les microbloggeurs peuvent catégoriser ou suivre des sujets à l'aide des hashtags.

b. Facteurs basés sur l'hypertextualité :

— Présence d'une URL dans le tweet : partager des URL est une manière de confirmer l'information publiée dans un tweet. Ceci permet également d'attirer l'attention sur un contenu présent sur le web. Ainsi, la présence d'une URL dans un tweet indique que le tweet a un caractère informatif renforcé.

— Fréquence de l'URL dans le corpus : ce critère permet de calculer la popularité des URL publiées dans un tweet dans le corpus.

c. Facteurs basés sur la popularité des auteurs :

— Nombre de tweets de l'auteur [Nagmoti et al., 10] : l'objectif de ce critère est de valoriser les tweets publiés par des auteurs actifs par rapport aux tweets publiés par des auteurs moins actifs. L'idée de ce critère est que les utilisateurs actifs ont plus de valeur en tant que sources d'information que des utilisateurs moins actifs.

— Nombre de références de l'auteur [Zhao et al., 11] : plus un auteur est référencé (ou mentionné), plus il est populaire.

La recherche d'information dans Twitter combine la recherche thématique (tweets traitant du sujet de la requête) et les facteurs de pertinence sociale, au sein d'un même score pour tenter de retrouver les tweets pertinents pour une requête donnée.

6.2.3 Évaluation de la recherche d'information dans les microblogs :

a. La tâche TREC Microblog :

C'est une tâche mise en place dans la campagne d'évaluation TREC. Depuis 2011, trois versions de cette tâche ont été mises en œuvre (2011, 2012 et 2013), offrant chacune une collection de testes correspondante.

— La collection de test Tweets2011 : comprend 16 millions de tweets (0,5 Go) exprimés dans diverses langues et publiés sur Twitter entre le 23 janvier 2011 et le 8 février 2011. Le corpus est conçu pour être un échantillon réutilisable et représentatif de la twittosphère.

— La collection de test Tweets2012 : comprend le même corpus de tweets que celui de 2011, avec 60 nouvelles requêtes avec leurs jugements de pertinence. Seuls les tweets hautement pertinents ont été considérés dans l'évaluation des systèmes.

— La collection de test Tweets2013 : comprend une nouvelle collection de 240 millions de tweets (70 Go), publiés dans la période du 1er février 2013 au 31 mars 2013. Cette collection est accessible uniquement à travers une API (contrairement à l'ancienne collection). Elle comprend 60 nouvelles requêtes avec les jugements de pertinence associés.

b. Les mesures d'évaluation utilisées dans TREC microblogs :

— F-mesure : c'est une mesure qui combine le rappel et la précision, elle est définie comme suit : $F - mesure = 2 * RP/(R + P)$

— La précision moyenne MAP : est utilisée comme une mesure supplémentaire pour évaluer l'efficacité de recherche, tout en tenant compte de la précision, du rappel et du rang des documents.

— La précision p@30 : est la mesure officielle pour l'évaluation de la tâche de recherche en temps réel dans TREC microblog 2011. Cette mesure évalue la capacité d'un système à retourner les tweets pertinents, parmi les 30 premiers de la liste des résultats.

Conclusion :

Dans ce chapitre nous avons défini en premier lieu les principales notions et concepts de la recherche d'information tels que le processus de la RI (indexation, appariement), les différents modèles de recherche d'information et le principe de l'évaluation des systèmes de RI. En deuxième lieu, nous avons défini la recherche d'information dans Twitter, nous avons introduit les facteurs de pertinence dans les microblogs, les différents aspects de la recherche d'information dans Twitter.

Dans le prochain chapitre, nous nous intéresserons à la recherche d'influenceurs et à la mesure de l'influence des bloggeurs sur Twitter.

Chapitre 02 :

Mesures d’influence sociale dans Twitter

Introduction :

L'analyse de l'influence sociale est un sujet émergeant dans le domaine de la recherche des réseaux sociaux en ligne. Alors que le microblog devient un média de masse d'une importance vitale, les mesures de l'influence sociale des utilisateurs de microblogging sont de plus en plus prises en compte. Dans ce chapitre, nous définissons les notions d'influence en général et d'influence sociale, puis l'influence sur twitter et quelques indicateurs de mesure d'influence sur twitter et ensuite nous exposons les travaux de mesure de l'influence dans Twitter.

1. L'influence sociale :

Le dictionnaire Larousse définit le mot **influence** comme : *le pouvoir social et politique de quelqu'un, d'un groupe, qui leur permet d'agir sur le cours des événements ou des décisions prises.*

L'influence [F.Boubekeur, 17] peut être définie comme la capacité d'une personne à s'imposer aux autres par la simple force de ses idées, et/ou de sa popularité, et/ou de sa notoriété. Un utilisateur d'un réseau social est influent si ses actions dans le réseau sont capables d'affecter les actions de nombreux autres utilisateurs du réseau [Fabian Riquelme, 16].

L'influence sociale est une relation établie entre une entité A dite "influenceur", et une autre entité B "l'influencé", qui a pour effet de provoquer chez B, une réaction, ou un comportement visés en réponse à une action de A [Anger, 11]. Dans (D.Cercel et S.trausan-Matu, 2014), les auteurs définissent l'influence sociale comme suit : soit A et B deux utilisateurs d'un réseau social donné, A a un pouvoir sur B, c'est-à-dire que A a la capacité de modifier l'opinion de B de manière directe ou indirecte.

Trois types d'influences sociales peuvent être distingués comme suit :

- le conformisme : c'est l'influence de la majorité sur l'individu ;
- l'innovation : définie comme l'influence qu'a un individu ou une minorité de personnes sur une majorité ;

- la soumission à l'autorité : définie en psychologie comme la réalisation d'une conduite prescrite par une source d'autorité (réalisation d'une obligation).

2. L'influence sur Twitter :

L'influence sur Twitter est devenue un sujet de recherche important, et a suscité l'intérêt de plusieurs experts qui ont proposé plusieurs approches différentes afin de mesurer cette influence. Dans ce qui suit, nous définissons les influenceurs dans les réseaux sociaux et dans twitter en particulier, les facteurs d'influence ainsi que les indicateurs de mesure de l'influence dans twitter, ensuite nous passons en revue la littérature des approches et mesures proposées auparavant.

2.1 Les influenceurs :

Le leader d'opinion (influenceur) est un individu qui par son autorité, son expertise ou son activité sociale intensive est susceptible d'influencer les opinions ou actions d'un grand nombre d'individus [Bertrand Bathelot, 17].

Les statuts des influenceurs peuvent prendre des formes variées : ils peuvent être des journalistes spécialisés, des blogueurs, des consultants ou encore des représentants associatifs. Ils se distinguent généralement par leur grand nombre d'abonnés mais aussi par le nombre de leurs publications, ils sont vraiment efficaces pour transmettre leurs opinions.

2.2 Définition d'un influenceur sur twitter :

Un twitto ou twitteur est une personne qui exprime une idée ou transmet une information en 280 caractères sur twitter. Il est dit «influent» s'il a la capacité de s'imposer sur le réseau social par ses idées et ses informations.

2.3 Les indicateurs de mesures de l'influence sur twitter :

Pour mesurer l'influence [Camille Jourdain, 09], cinq indicateurs principaux ont été pris en considération :

- **Le nombre de followers :** indicateur d'audience ;
- **Le nombre de followers des followers :** indicateur d'audience ;
- **Le nombre de followings des followers** : indicateur de visibilité ;

— **Le rapport entre le nombre de followings et le nombre de followers** : indicateur de réputation ;

— **Le rapport entre le nombre de citations et le nombre de tweets :** indicateur d'écoute et de reconnaissance.

Pour mesurer l'influence d'un Twitto, différentes approches ont été proposées. Dans ce qui suit, nous introduisons une revue de la littérature sur ce sujet.

3. Approches de mesure de l'influence sur Twitter :

3.1 Mesure traditionnelle :

Les mesures d'influence de Twitter de la littérature prennent généralement en compte les mesures liées aux relations de retweets, mentions et même d'abonnement. Cependant, certains chercheurs ont utilisé les mesures de centralité qui sont basées sur la topologie du réseau social. C'est le cas des mesures de *Closeness* (Cc) et de *Betweenness* (CB). [Fabian Riquelem, 16].

— Closeness Centrality : Cette mesure est basée sur la longueur du chemin le plus court d'un nœud i vers tous les autres nœuds. Elle mesure la visibilité ou l'accessibilité de chaque nœud par rapport à l'ensemble du réseau. Soit D la matrice de distance d'un réseau avec n nœuds, l'équation est définie comme suit :

$$Cc = \frac{n-1}{\sum_{i \neq j} (D)ij}$$

— Betweenness Centrality : Cette mesure considère pour un nœud i, tous les chemins les plus courts qui doivent passer par i pour connecter tous les autres nœuds du réseau. Elle mesure la capacité de chaque nœud à faciliter la communication au sein du réseau. Soit b_{jk} le nombre des chemins les plus courts du nœud j au nœud k, et b_{jik} le nombre de ces chemins les plus courts qui traversent le nœud i, la mesure de betweeness est définie comme suit :

$$Cb = \frac{1}{(n-1)(n-2)} \sum_{i \neq j} \sum_{k \neq j \text{ et } k \neq i} \frac{b_{jik}}{b_{jk}}$$

3.2 L'approche Tunklang :

Dans [Tunklang, 09], l'auteur propose une mesure de l'influence d'un blogueur calculée sur la base du nombre de ses abonnés (followers) et du nombre de ses amis (followings). Partant de l'idée qu'un utilisateur est d'autant plus influent qu'il est suivi par d'autres utilisateurs influents, et que lui-même ne suit pas beaucoup d'autres utilisateurs du réseau, l'auteur propose d'estimer l'influence d'un blogueur comme suit :

$$Influence(u) = \frac{followers}{followings}$$

Tel que :

- Followers : est le nombre de ses abonnés ;
- Followings : est le nombre de ses amis.

3.3 L'approche Cha et al :

Les auteurs Cha et al. Ont défini en 2010 [Cha, 10] trois mesures de l'influence d'un utilisateur dans le réseau social de Twitter :

— L'influence indue (*Indegree influence*) : mesure le nombre de followers d'un utilisateur ; elle indique directement la taille de l'audience pour cet utilisateur u.

— L'influence du retweet (*Retweet influence*) : mesure le nombre de fois que d'autres utilisateurs ont rediffusé les tweets publiés par un utilisateur.

— L'influence de la mention (*Mention influence*) : mesure l'influence d'un utilisateur u à travers le nombre de mentions contenant son nom et indique la capacité de cet utilisateur à engager d'autres personnes dans une conversation.

Dans cette approche, les auteurs ont constaté que les retweets et les mentions sont corrélés à l'influence et sont donc de bons indicateurs de l'influence, contrairement au nombre d'abonnés (à travers *indegree*) qui représente la popularité d'un utilisateur et révèle très peu sur l'influence d'un utilisateur.

3.4 L'approche Weng et al :

Dans [Weng, 10], les auteurs ont été les premiers à signaler le phénomène de l'homophilie dans une communauté de Twitter, ainsi que dans le contexte de Twitter. L'homophilie implique qu'un twitto suit un ami parce qu'il s'intéresse à certains sujets que l'ami publie, et l'ami le suit en retour parce qu'il trouve qu'ils partagent un intérêt d'actualité similaire.

Les auteurs ont noté qu'il existait une forte réciprocité entre blogueurs sur le réseau social d'abonnement ; Cela est dû à leur observation que 72,4% des utilisateurs suivent plus de 80% de leurs abonnés et 80,5% des utilisateurs ont 80% de leurs amis qui les suivent. Ainsi les auteurs ont déduit que deux raisons contradictoires peuvent expliquer une telle réciprocité. Soit la relation est juste au hasard car un utilisateur suit quelqu'un par souci de courtoisie. Ou bien, cette relation peut être une preuve de similarité, autrement dit, un twitto suit un ami parce qu'il s'intéresse à ses publications et vice-versa.

En se basant sur cette observation, les auteurs ont proposé une approche pour mesurer l'influence sur twitter, nommée TwitterRank, une extension de PageRank[1] qui exploite la structure du réseau social basé sur les relations d'abonnements. Cet algorithme se distingue de PageRank dans le sens que l'utilisateur fait une recherche spécifique à un sujet c'est-à-dire, la probabilité de transition d'un twitto à un autre est spécifique à un sujet, formellement :

$$\mathrm{P(i,\,i)} = \frac{|T_j|}{\sum_{\propto:\, i\, follows\, j} * T_{\propto}} * sim_t(i,j)$$

Tel que :

- $|T_j|$: est le nombre de tweets publiés par j ;
- $\sum_{\propto:\, i\, follows\, j} * T_{\propto}$: Le nombre de tweets publiés par tous les amis de i ;
- $sim_t(i,j)$: La similarité entre i et j sur un sujet t.

[1] L'algorithme PageRank développé en 1998 est à l'origine du moteur de recherche Google. Cet algorithme assigne un score à toutes les pages du web indépendamment de toute requête. Quant à la mesure PageRank, c'est une distribution de probabilité sur les pages. Elle mesure la probabilité pour un utilisateur navigant au hasard, d'atteindre une page donnée. Elle repose sur le concept qu'un lien émis par une page A vers une page B est assimilé à un vote de A pour B. Plus une page reçoit de votes, plus elle est considérée importante.

3.5 L'approche Romero et al :

Les auteurs dans [Romero D. M, 11] ont proposé l'algorithme IP (*Influence-Passivity algorithm*) inspiré de l'algorithme HITS[2] [Kleinberg. 99] qui détermine l'influence et la passivité des utilisateurs en fonction de leur activité de transfert d'informations. Dans leur étude, les auteurs ont montré que la majorité des utilisateurs agissent comme des consommateurs passifs d'informations et ne transmettent pas le contenu au réseau (un utilisateur passif est un utilisateur qui ne retweet pas (ou peu) les tweets des autres). Par conséquent, pour que les individus deviennent influents, ils doivent non seulement attirer l'attention et donc être populaires, mais aussi surmonter la passivité des utilisateurs.

Cet algorithme (IP) attribue un score d'influence relatif et un score de passivité à chaque utilisateur. La passivité d'un utilisateur est une mesure de la difficulté pour d'autres utilisateurs à l'influencer.

3.6 L'approche Anger et Kittl :

Dans leur approche [Anger, 11], les auteurs Anger et Kittl se sont focalisés sur les relations d'abonnement, de retweet et de mention, et ont défini les paramètres de mesure d'influence suivants :

- Le ratio d'abonnement : définit comme le nombre d'abonnés divisé par 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.
- Le ratio d'interaction : définit par le nombre d'utilisateurs qui retweetent ou mentionnent divisé par le nombre d'abonnés.

Les auteurs suggèrent que le ratio de followers (ou d'abonnés) utilisé seul, n'est pas un bon indicateur d'influence car d'une part de nombreux utilisateurs suivent d'autres utilisateurs pour gagner un suivi mutuel, d'autre part, il existe plusieurs entreprises qui offrent des clics de suivi pour une certaine somme d'argent. Donc, ce sont principalement les interactions et les réactions positives qui déterminent le succès d'un utilisateur. Partant de là, les auteurs

[2] Le modèle de Kleinberg à la différence de PageRank distingue les « autorités » (pages recevant beaucoup de liens) des « Hubs » (pages contenant beaucoup de liens vers de bonnes pages). Le moteur de recherche ask.com est basé sur l'algorithme HITS.

proposent une mesure de l'influence dite **potentiel de réseautage social** SNP (*Social Networking Potential*), calculée comme la moyenne du ratio de retweet et d'interaction.

3.7 L'approche Ben Jabeur et al :

Les auteurs [Benjabeur L, 11] ont proposé de modéliser le réseau social de Twitter en se basant sur les relations de rediffusion. L'importance d'un blogueur est alors déterminée par la proportion de ses messages rediffusés. Le réseau social d'influence est modélisé par un graphe G = (U, E) où U est l'ensemble des utilisateurs et E = U × U représente l'ensemble des relations d'influence entre eux. Une relation d'influence e (ui, uj) ϵ E est définie si et seulement s'il existe au moins un article publié par uj et rediffusé par ui. Le poids w (ui, uj) de la relation d'influence est calculé par la formule suivante :

$$w(ui, uj) = \frac{\text{nombre d}'\text{ articles publiés par uj et rediffusés par ui}}{\text{nombre d}'\text{ articles rediffusés par ui}}$$

3.8 L'approche Ding et al :

Dans [Ding, 13], les auteurs ont proposé un algorithme similaire à PageRank, nommé SpreadRank qui mesure l'influence d'un utilisateur à travers sa propagabilité (*spreadability*), en se basant sur le réseau de rediffusion des tweets. La propagabilité d'un utilisateur est sa capacité à propager l'information sur le réseau social. Les auteurs l'ont mesuré sur une cascade d'information dont les arcs sont pondérés par le rapport retweets/nombre total de tweets publiés par l'utilisateur retweeté. La figure 2.1 illustre un exemple d'une cascade d'information de profondeur quatre.

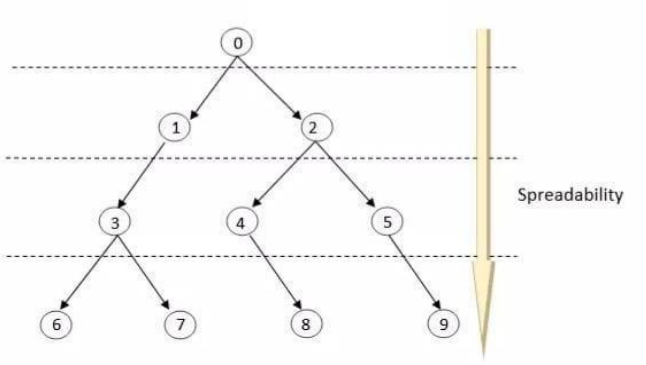


Figure 2.1 : Exemple de cascade d'information

- La propagation des nœuds supérieurs est plus élevée en comparaison avec celle des nœuds inférieurs.

L'algorithme SpreadRank tient en compte la localisation des utilisateurs dans la cascade d'information ainsi que l'intervalle de temps entre les retweets. Les auteurs ont combiné ces deux mesures pour calculer la probabilité de transition de chaque utilisateur.

3.9 L'approche de Sung et al :

Les auteurs [Sung J, 13] ont proposé un algorithme similaire à pageRank, intitulé InterRank (*Interaction Rank*), qui exploite les relations d'abonnements dans le réseau social twitter. Les auteurs ont confirmé que les gens préféraient interagir avec des personnes similaires (même sujet d'intérêt). De même que cette similarité est liée à la diffusion d'influence dans Twitter. Formellement :

$$\text{InterRank (pi)} = \frac{1-\text{d}}{N} + \text{d} \sum_{pi \in M(pi)} CS(pi, pj) * \frac{\text{InterRank(pj)}}{\text{L(pj)}}$$

Ou :

- Pi : représente un utilisateur de twitter ;
- M (pi) : représente l'ensemble des suiveurs qui suivent un utilisateur pi ;
- L (pj) : le nombre d'amis d'un utilisateur twitter pj ;
- CS (pi, pj) : représente le cosine similarity entre les utilisateurs pi et pj ;
- d : est un paramètre généralement fixé à 0.85 ;
- N : nombre de nœuds du réseau.

3.10 L'approche Zhang et al :

Dans leur approche, les auteurs [Zhang M, 11] ont utilisé un réseau basé sur l'action pour mesurer l'influence. En particulier deux types d'actions utilisateur sont considérés : le retweet et la réponse. Les deux actions sont destinées à transmettre le contenu à d'autres utilisateurs de différentes manières.

— Si un utilisateur répond au tweet d'un autre utilisateur (action de réponse), cela implique que l'utilisateur est influencé par le tweet.

— Si un utilisateur cite ou paraphrase le contenu d'un autre utilisateur (action de retweet), cela implique que l'utilisateur est influencé pour reproduire le contenu.

L'idée est que, si un utilisateur influent retweete ou répond à un autre utilisateur, alors ce dernier est probablement influent. En outre, plus un utilisateur retweete ou répond à d'autres utilisateurs, plus son influence se propage. Les auteurs modélisent la propagation d'influence sur les réseaux de retweets (retweet network) et de réponses (Reply network) de Twitter respectivement, en utilisant l'algorithme PageRank.

3.11 L'approche Azaza et al :

Les auteurs dans [Azaza et al, 15] ont proposé une nouvelle mesure de l'influence multicritère qui consiste à combiner sept critères d'influence, à savoir : les mentions, les retweets, les hashtags, le lien URL, les réponses, les followers et les favoris.

3.12 L'approche Kwak et al :

Les auteurs [Haiwoon Kwak, 10] ont comparé trois mesures d'influence différentes : le nombre d'abonnés, pageRank et le nombre de retweets, ils ont fini par déduire que le classement des utilisateurs les plus influents, différait selon la mesure.

Pour ce fait, ils ont considéré 20 célébrités de différents domaines (acteurs, musiciens, politiciens…), ils ont déduit que :

— Le nombre de followers favorise plus les acteurs, musiciens et les animateurs TV ;

— PageRank favorise : les acteurs, présidents, et news ;

— Le nombre de retweets favorise les news.

3.13 F.Boubekeur, M.Ferrouk, L.Belkacemi :

Ici les auteurs [Boubekeur et al., 17] ont proposé une mesure de l'influence d'un blogueur sur Twitter à travers un ratio d'influence. Le ratio d'influence d'un utilisateur est défini comme le rapport entre :

— Sa capacité à influencer d'autres blogueurs (influence imposée),

— Et sa disposition à être influencé par d'autres blogueurs (influence subie).

Le ratio d'influence est calculé sur un réseau d'influence basé sur les retweets, par un algorithme de propagation adapté de PageRank. Formellement :

$$\tau\text{inf}\,(\text{ui}) = \frac{\text{Infl\_i}(u_i) + 1}{\text{Infl\_s}(u_i) + 1} + ?$$

Tel que :

- τinf (ui) : est le ratio d'influence du blogueur ui dans le réseau social d'influence ;
- μ = 0.05 : constante utilisée pour assurer la convergence du calcul récursif du ratio d'influence ;
- le " + 1 "dans la formule évite la division par zéro ;
- Infl_i (ui) : influence imposée par ui sur les autres blogueurs du réseau social d'influence.

Formellement :

$$\text{Infl\_i(ui)} = \sum_{u_j\,,u_j\,retweet\,u_i} w(u_i, u_j) * \tau_{inf}(u_j)$$

Tel que :

$$w(u_i, u_j) = \frac{\text{nombre de tweets publies par } u_i \text{ et rediffuses par } u_j}{\text{nombre de tweets publies par } u_i}$$

- $\tau_{inf}(u_j)$: est le ratio d'influence du blogueur uj dans le réseau social d'influence ;
- Infl_s (u_i) : influence subie par le blogueur ui sur le réseau social d'influence.

Formellement :

$$\text{Infl\_s(ui)} = \sum_{u_k\,,u_i\,retweet\,u_k} w(u_k, u_i) * \tau_{inf}(u_k)$$

Tel que :

$$w(u_{k}, u_i) = \frac{\text{nombre de tweets publies par } u_k \text{ et rediffuses par } u_i}{\text{nombre de tweets publies par } u_k}$$

- $\tau_{inf}(u_k)$: est le ratio d'influence du blogueur u_k dans le réseau social d'influence.

Conclusion :

Dans ce chapitre nous avons pu voir l'influence sociale sur les plateformes de microblogging. Par la suite, nous nous sommes basées sur l'influence sur Twitter en exposant plusieurs études phares en relation avec cette thématique.

Le prochain chapitre portera sur l'approche que nous proposons pour la mesure de l'influence sur twitter.

Chapitre 03 :

L’approche proposée

Introduction :

Dans ce chapitre, nous présentons notre contribution à la définition d'une nouvelle mesure d'influence que nous utilisons pour définir un nouveau modèle de RI social.

1. Description de l'approche proposée :

1.1 Modélisation du réseau social Twitter :

Le réseau social se constitue de différents utilisateurs et d'un ensemble de relations (retweet, replay, mention, abonnement …) qui les lient. Ce réseau social est modélisé par un graphe G= (U, E) où U et E=U x U représentent respectivement l'ensemble des twittos et les relations d'influence entre eux. Dans notre travail nous nous intéressons à la relation de retweets.

1.2 Mesure de l'influence d'un Twitto :

Un influenceur n'est pas nécessairement celui qui a la plus forte audience. L'influence se mesure surtout par la capacité à engager et à faire réagir les autres. Plusieurs approches ont été passées en revue en chapitre 2 qui proposent de mesurer l'influence d'un Twitto en se basant sur différents signaux sociaux issus des relations sociales sur Twitter.

Pour notre part, nous proposons d'utiliser la relation de retweet. L'intuition derrière l'utilisation de cette relation vient de l'observation que :

— Si un twitto est influent alors la plupart de ses tweets sont retweetés.

—Si les tweets d'un blogueur sont fortement retweetés (la somme de retweet est importante), ce blogueur devient par ce fait influent.

1.2.1 Le ratio de retweet :

$$Ra_{retweet}(u) = \frac{nombre\ de\ tweets\ retweeté(u) + 1}{nombre\ de\ tweets\ publié(u) + 1}$$

Où :

— 1 est une valeur de lissage qui évite une division par zéro.

1.2.2 Le ratio de la somme de retweet :

$$Ra_{\sum retweet}(u) = \frac{\sum retweet\ (u) + 1}{nombre\ de\ tweets\ publié + 1}$$

Où :

— 1 est une valeur de lissage qui évite une division par zéro.

1.2.3 Le H_index :

Le h_index (facteur h) est un indicateur bibliométrique utilisé pour donner l'impact d'un chercheur, créé par le physicien Jorge Hirsch en 2005. C'est un indicateur d'impact des publications d'un chercheur. Il prend en compte le nombre de publications d'un chercheur et le nombre de leurs citations.

L'indice h d'un auteur est égal au nombre h le plus élevé de ses publications qui ont reçu au moins h citations chacune. Il est calculé en classant et en numérotant les publications de l'auteur de la plus citée (n° 1) à la moins citée. L'indice h correspond au dernier numéro de la publication qui vérifie :

$$numéro\ de\ la\ publication \leq nombre\ de\ citations$$

Dans notre cas, nous pouvons calculer l'indice 'h' en classant et en numérotant les tweets d'un blogueur par ordre décroissant selon leurs nombre de retweets. L'indice h correspond au dernier numéro du tweet qui vérifie :

$$numéro\ du\ tweet \leq nombre\ de\ retweets$$

1.2.4 Mesure d'influence combinée :

Nous combinons le h-index et les ratios précédents en un score d'influence comme suit :

$$Infl(u) = Ra_{retweet}(u) * Ra_{\sum retweet}(u) * h\_index(u)$$

Où :

— Infl(u) : représente le score d'influence d'un Twitto.

1.3 Exemples explicatifs :

Dans cette section, nous présentons trois exemples explicatifs pour illustrer la pertinence de notre score d'influence. Dans les exemples suivants les tweets sont classés par ordres décroissants selon leur nombre de retweets pour chacun des twittos.

1.3.1 Exemple 1 :

Voici deux twittos, Twitto 1 et Twitto 2 qui ont deux tweets chacun, comme le montre le tableau 3.1 suivant :

Tableau 3.1 : Application de la loi de calcule

	Tweets	**Retweet**	**∑retweet**	**h-index**	**$Ra_{retweet}$**	**$Ra_{\sum retweet}$**	**l'influence**
Twitto 1	**t_1**	100	100	1	0,66	33,66	22,21
	t_2	0					
Twitto 2	**t_1**	1	1	1	0,66	0,66	0,43
	t_2	0					

Les deux twittos ont le même nombre de tweets et même nombre de tweets retweetés, d'où, ils obtiennent un ratio de retweets équivalents. Cependant, ils diffèrent par la somme de retweets, ce qui a abouti à la domination du Twitto 1 par sa somme de retweets élevée.

Le score d'influence de Twitto 1 est supérieur à celui du Twitto 2.

1.3.2 Exemple 2 :

Voici quatre twittos : Twitto 1, Twitto 2, Twitto 3 et Twitto 4 qui ont le même nombre de tweets publiés et la même somme de retweets, comme le montre le tableau suivant :

Tableau 3.2 : Application de la loi de calcule

	Tweets	**Retweet**	**∑retweet**	**h-index**	**$Ra_{retweet}$**	**$Ra_{\sum retweet}$**	**l'influence**
Twitto 1	t_1	20	50	3	1	10,2	30,6
	t_2	15					
	t_3	14					
	t_4	1					
Twitto 2	t_1	18	50	3	0,8	10,2	24,4
	t_2	17					
	t_3	15					
	t_4	0					
Twitto 3	t_1	27	50	2	0,6	10,2	12,24
	t_2	23					
	t_3	0					
	t_4	0					
Twitto 4	t_1	50	50	1	0,4	10,2	4,08
	t_2	0					
	t_3	0					
	t_4	0					

Bien que les quatre Twittos aient la même somme de retweets et malgré l'équivalence de h_index des deux premiers Twittos, nous observons que le score d'influence du Twitto 1 est

supérieur à ceux des trois autres Twittos. Cela revient à son ratio de retweets élevé par rapport aux ratios des autres Twittos vu que tous ses tweets ont été retweetés.

1.3.3 Exemple 3 :

Voici deux Twittos Twitto 1 et Twitto 2 qui ont le même nombre de tweets et même nombre de tweets retweetés. Cependant la somme des retweets et le h_index sont différents comme le montre le tableau ci-dessous :

Tableau 3.3 : Application de la loi de calcule

	Tweets	**Retweet**	**∑retweet**	**h_index**	**$Ra_{retweet}$**	**$Ra_{\sum retweet}$**	**l'influence**
Twitto 1	t_1	50	100	3	1	20,2	60,6
	t_2	25					
	t_3	24					
	t_4	1					
Twitto 2	t_1	20	25	2	1	5,2	10,4
	t_2	3					
	t_3	1					
	t_4	1					

Bien que les deux Twittos aient le même ratio de retweets, nous remarquons que le Twitto 1 qui a une somme de retweets et un h_index plus élevé a obtenu un score d'influence meilleur par rapport au Twitto 2.

— Observation :

En parcourant les trois exemples présentés ci-dessus, nous observons que notre score d'influence favorise les Twittos qui ont à la fois un ratio de retweets et un h_index élevés ainsi une somme de retweets importante.

2. Un nouveau modèle de recherche d'information social dans Twitter :

Dans la recherche d'informations, le but est de trouver rapidement une information pertinente par rapport à un objectif précis. Nous proposons d'utiliser notre mesure d'influence dans un contexte de recherche d'information sociale dans Twitter qui permet de tenir compte de la pertinence thématique et de la pertinence sociale du tweet par rapport à une requête. La pertinence thématique d'un tweet peut être calculée par tout modèle de la RI. Cette pertinence représente le degré de similarité entre le tweet et la requête. Par contre la pertinence sociale d'un tweet est liée à l'influence de son auteur mesurée par notre score d'influence comme défini plus haut (Infl(u)).

Nous définissons un modèle de recherche social qui combine la pertinence thématique d'un tweet à sa pertinence sociale dans un score linéaire comme suit :

$$Rel(Q,t) = \alpha\ RSV\ (Q,t) + (1-\alpha) Infl(u)$$

Où :

- RSV (Q, t) : est la pertinence thématique du tweet t pour la requête Q. Elle est calculée comme le degré de correspondance du tweet pour la requête ;
- α : est un paramètre compris entre 0 et 1.

3. Conception d'une solution de RI basée sur le facteur d'influence :

Afin de modéliser les déférents aspects de notre application, nous avons opté pour l'emploi du langage de modélisation unifié UML. La notation UML est un langage visuel constitué d'un ensemble de schémas, appelés « diagrammes », qui donnent chacun une vision différente du projet à traiter.

Nous avons utilisé le diagramme de cas d'utilisation pour donner une vision globale du comportement fonctionnel de notre système.

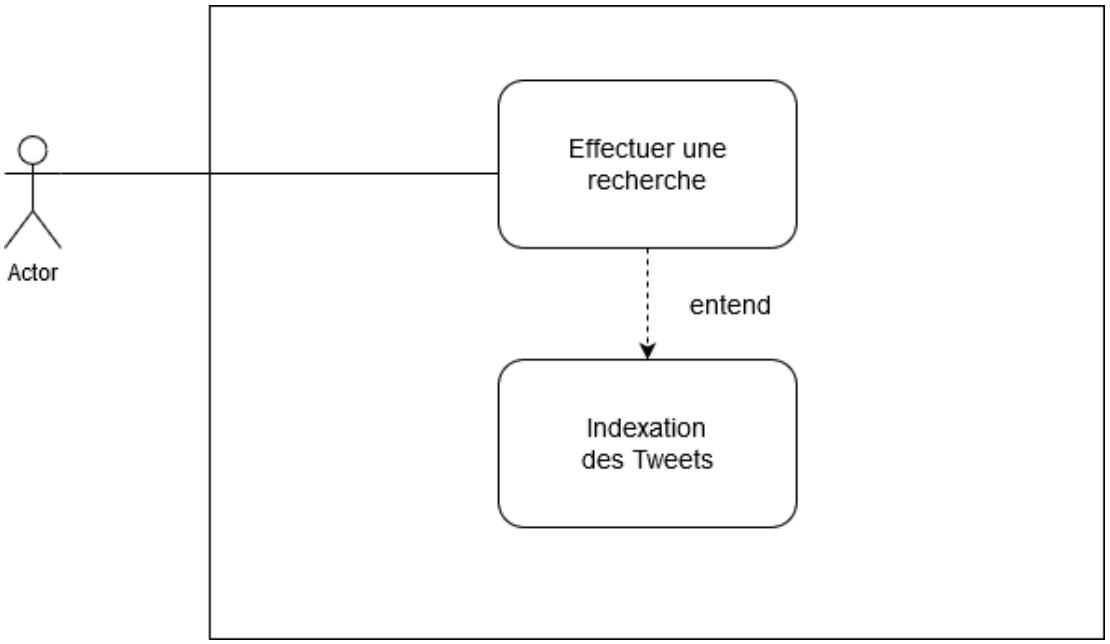


Figure 3.1 : Diagramme de cas d'utilisation

Nous avons subdivisé notre application en deux modules principaux : Indexation et Recherche qui sont respectivement consacré à l'indexation des documents et à l'exécution des recherches. Vu que les deux phases basiques pour toute application de recherche d'information sont l'indexation et la recherche.

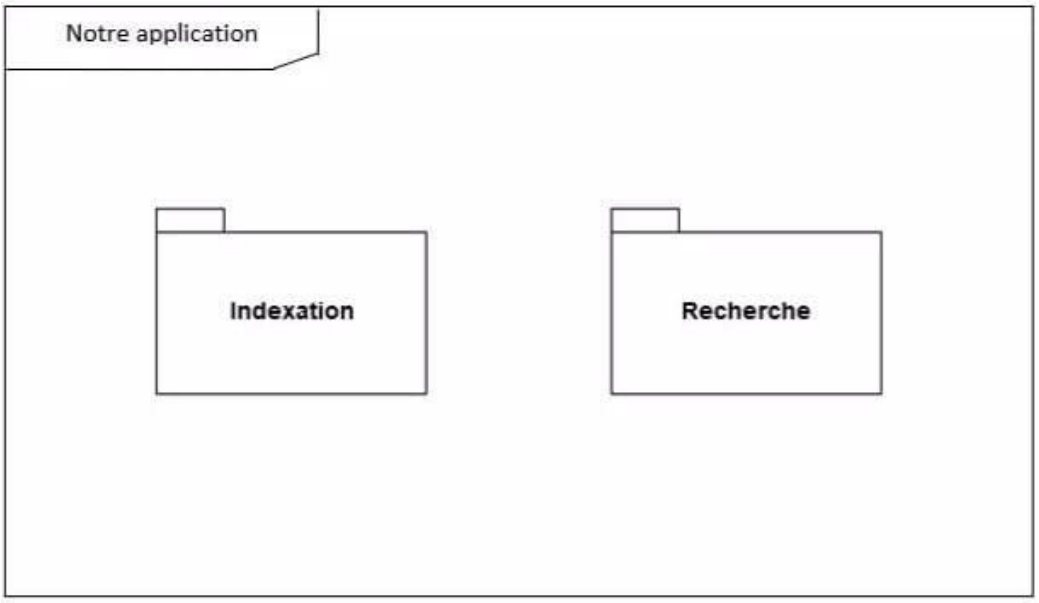


Figure 3.2 : Diagramme de package de l'implémentation

Pour la représentation de différents traitements effectué par notre système, nous avons utilisé le diagramme de séquence qui permet de représenter graphiquement les communications (la chronologie des échanges de messages) avec et au sein d'une application.

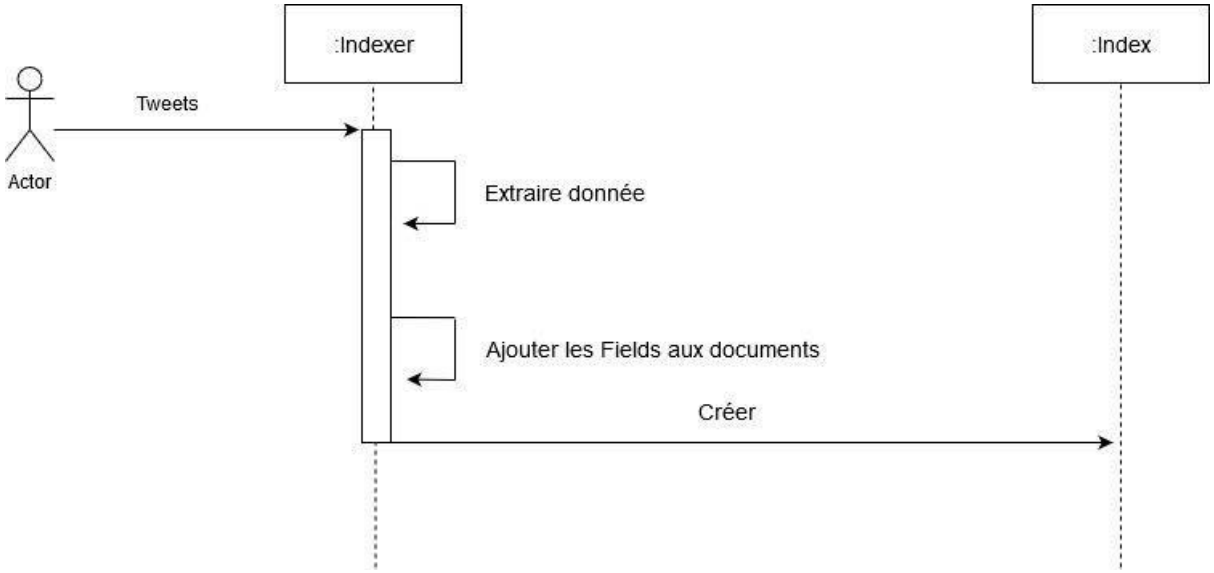


Figure 3.3 : Diagramme de séquence de l'indexation

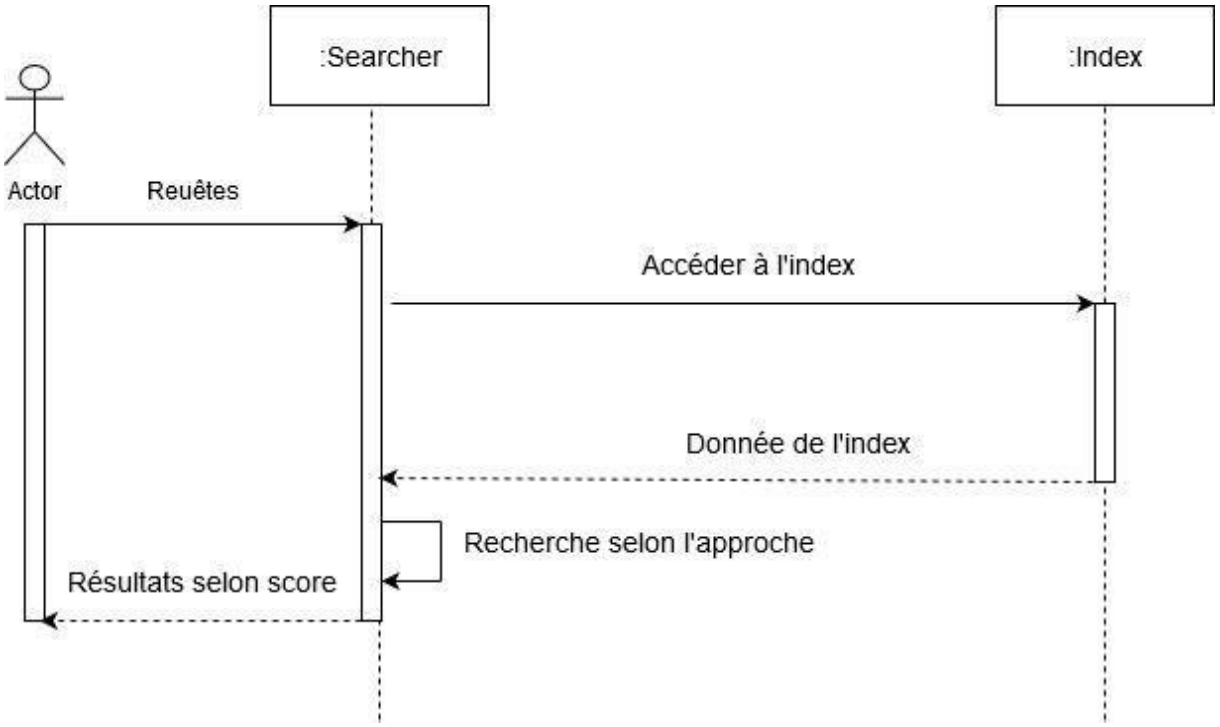


Figure 3.4 : Diagramme de séquence de la recherche

Conclusion :

Dans ce chapitre, nous avons décrit dans un premier temps notre approche de mesure de l'influence qui combine les deux ratios de retweet et de somme de retweet d'un twitto avec son h_index, par la suite nous avons présenté un aperçu de notre conception. Dans le chapitre suivant, nous allons expérimenter notre approche afin de voir si elle apporte une amélioration par rapport aux résultats obtenus par le score thématique.

Chapitre 04 :

Implémentation et évaluation

Introduction :

Dans ce chapitre, nous présentons les différents outils utilisés pour la réalisation de notre implémentation. Par la suite, nous présentons un aperçu de cette dernière.

1. Outils de développement :

1.1 Eclipse IDE :

Est un Environnement de développement permettant potentiellement de créer des projets de développement mettant en œuvre un langage de programmation. Eclipse IDE est principalement écrit en Java.

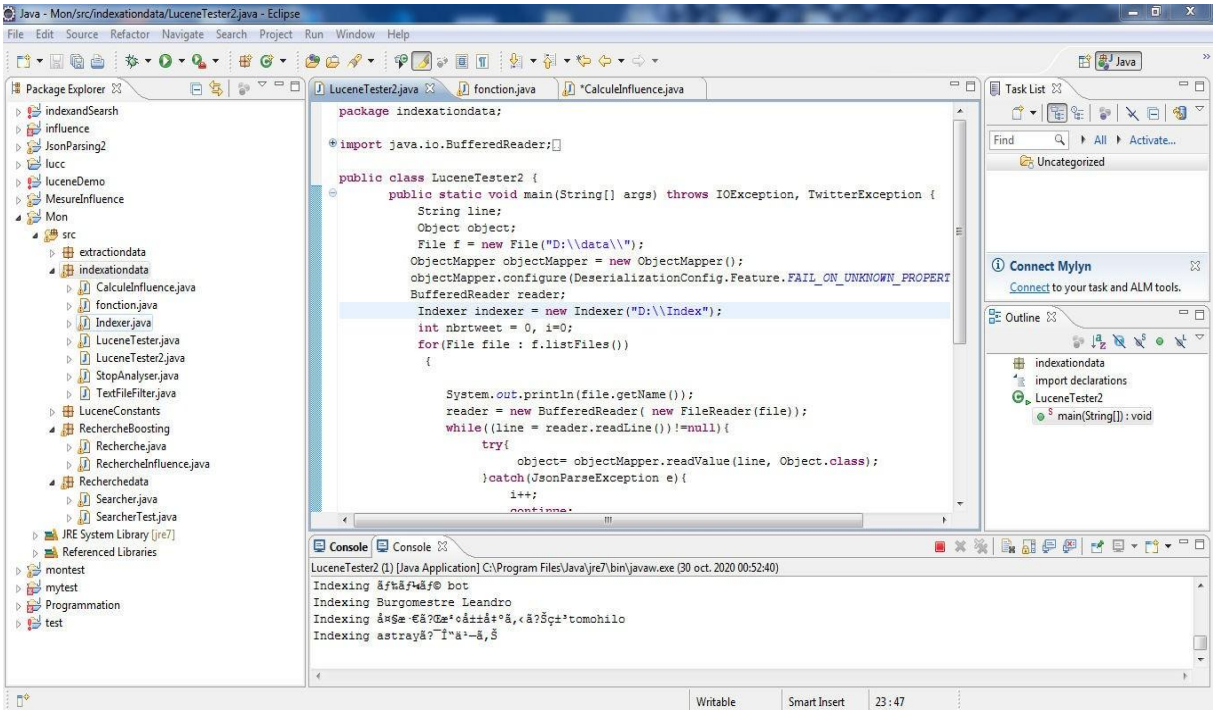


Figure 4.1 : Interface Eclipse IDE

1.2 Langage Java :

C'est un langage de programmation orienté objet, développé par Sun Microsystems. Il permet de créer des logiciels portables sur plusieurs systèmes d'exploitation (Windows, Linux,…). Il est rapide, sécurisé et fiable.

1.3 L'API Jackson :

JSON (JavaScript Object Notation) est un format (une syntaxe) standard utilisé pour représenter des données structurées (sérialiser des objets, tableaux, nombres, chaînes de caractères, booléens et valeurs nulles). Jackson est l'une des bibliothèques Java les plus populaires, permettant de désérialiser le flux JSON en objet métier ou inversement transformer (dit sérialiser) un objet métier en flux JSON.

1.4 Lucene 3.6 :

Lucene est une bibliothèque de recherche Java créée par Doug Cutting et développée par la fondation Apache. Elle peut être utilisée dans n'importe quelle application pour y ajouter une fonction de recherche. Cette bibliothèque haute performance est utilisée pour indexer et rechercher pratiquement tout type de texte (documents Word et PDF, e-mails, pages Web, tweets, etc.).

1.4.1 L'architecture de Lucene :

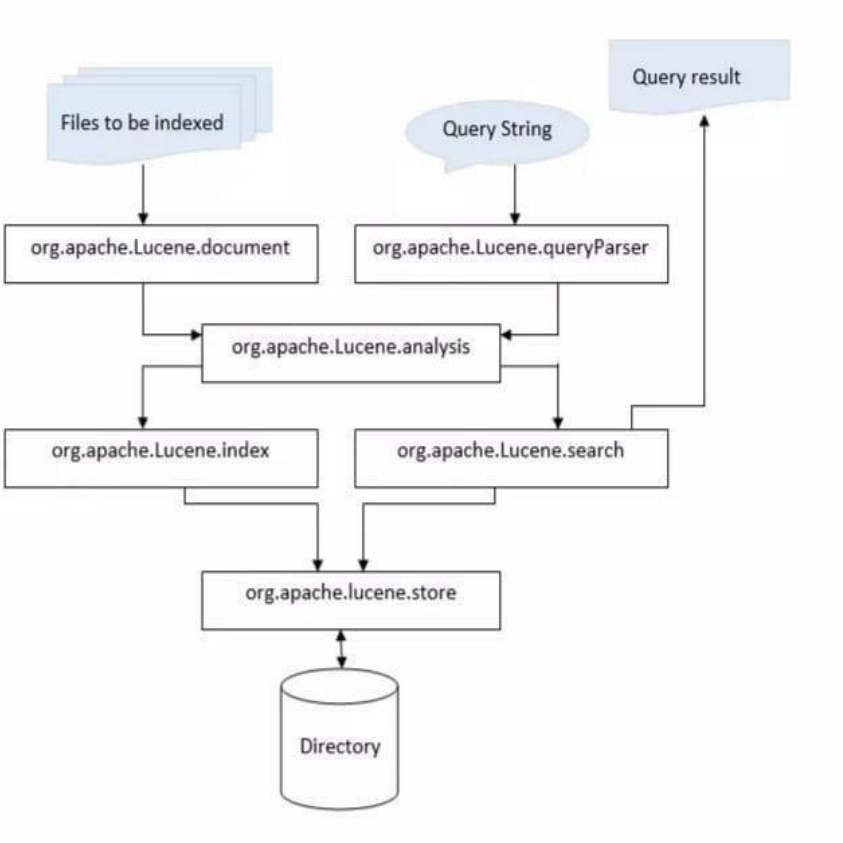


Figure 4.2 : Architecture de Lucene

Lucene se découpe en 7 Paquetages principaux qu'on retrouve dans l'architecture figure 4.2 :

- org.apache.lucene.analysis : Il contient le code pour convertir du texte en élément indexable. Il contient la classe Analyzer qui permet d'extraire les mots importants pour l'index et supprimer le reste.
- org.apache.lucene.document : Contient des classes relatives aux documents comme la classe Document qui représente un rassemblement de champs (fields), ainsi les métadonnées sont indexées et stockées séparément comme des champs d'un document.
- org.apache.lucene.index : Il contient le code pour accéder aux index. On y trouve la classe IndexWriter qui permet la création d'un nouvel index et l'ajout de documents à un index existant.

- org.apache.lucene.queryparser : les classes de ce paquetage tel que QueryParser ont pour responsabilité de parser (d'analyser) les requêtes pour générer la requête sous forme d'objet query qui pourront ensuite être réutilisés par le parseur.
- org.apache.lucene.search : se charge de fournir les objets pour chercher dans les indexes. Il fournit les classes IndexSearcher, Query et Hits :
  - IndexSearcher : permet de faire des recherches sur un index. Elle se charge de l'ouverture de l'index en lecture seule.
  - Query : est une classe abstraite, et le parent de tous les types de requêtes que Lucene utilise pendant le processus de recherche. Elle est utilisée pour égaliser les documents qui contiennent des champs avec des valeurs spécifiques.
  - Hits : La classe Hits est un simple conteneur d'index pour classer les résultats de recherche de documents qui apparaissent pour une interrogation donnée. Pour des raisons de performances, les exemples de classement ne chargent pas depuis l'index tous les documents pour une requête donnée, mais seulement une partie d'entre eux.
- org.apache.lucene.store : couche d'abstraction d'entrée sortie. On y trouve les classes suivantes :
  - Directory : est une classe abstraite, qui permet à ses sous-classes de stocker l'index à l'endroit désiré. Les fichiers peuvent être écrits une fois, lorsqu'ils sont créés. Une fois qu'un fichier est créé, il ne peut être ouvert qu'en lecture ou supprimé. L'accès aléatoire est autorisé à la fois en lecture et en écriture.
  - FSDirectory : elle étend de la classe Directory, elle sert à stocker des fichiers d'index dans le système de fichiers. Lucene comprend un certain nombre d'implémentations intéressantes de la classe Directory. Par exemple, l'implémentation FSDirectory.Open permet de stocker les fichiers dans un répertoire sur le système de fichiers.
- org.apache.lucene.util : classes utilisées dans les autres paquetages : une implémentation de tableau, une implémentation de vecteur de bits, constante concernant l'os utilisé, etc.

1.4.2 La recherche sous Lucene :

Le cœur d'un moteur de recherche basé sur Lucene est l'index. La fonction de recherche doit être précédée par la phase d'indexation. Le processus d'indexation et les classes utilisées sont illustrés dans la figure 4.3.

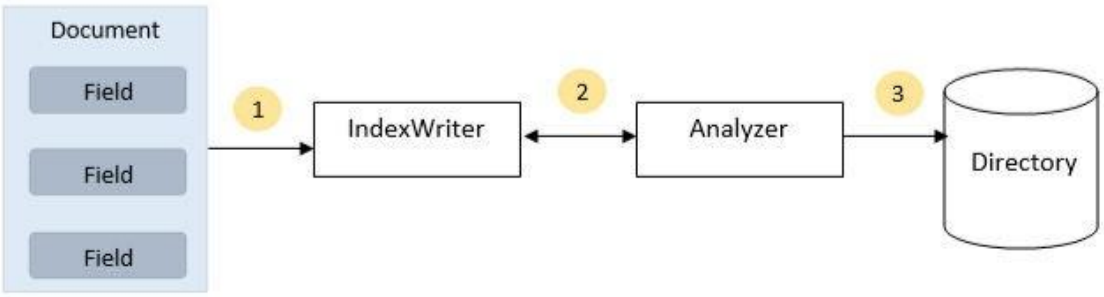


Figure 4.3 : Processus d'indexation

Nous ajoutons le ou les documents contenant le ou les champs à IndexWriter qui analyse le ou les documents à l'aide de l'Analyzer, puis nous créons, ouvrons ou éditons les index selon les besoins et nous les stockons ou mettons à jour dans un répertoire (Directory).

- Lucene indexe des objets appelés « Documents ». Un Document est une structure de données constituée de champs (field). Les champs d'un document représentent le document et ses métadonnées.
- La classe « Field » contient un nom (titre, auteur, date de publication, contenu, etc.) et une valeur généralement du texte qui est indexé, recherché et affiché.
- « IndexWriter » est la classe principale de la phase d'indexation, elle permet de créer un nouvel index (ou ouvrir un index existant), ajouter, supprimer ou mettre à jour les documents dans un index.
- « Analyzer » est un ensemble de classes ayant pour but le découpage du texte en token (mot) et la normalisation du texte à indexer.
- **«** Directory **»** représente l'emplacement de l'index de Lucene.

Lucene permet ensuite, à partir de l'index créé, une recherche rapide et efficace dans ces documents.

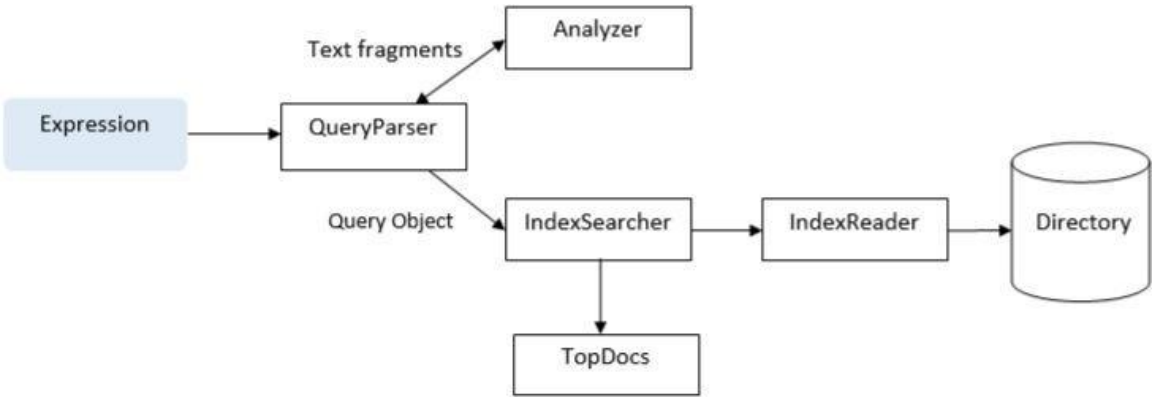


Figure 4.4 : Processus de Recherche

Une fois les répertoires contenant les index créés, nous les transmettons à **IndexSearcher** qui ouvre le répertoire à l'aide d'**IndexReader**. Ensuite, nous créons une requête avec un terme et effectuons une recherche à l'aide d'IndexSearcher en transmettant la requête au chercheur. IndexSearcher renvoie un objet TopDocs qui contient les détails de la recherche ainsi que les ID de documents qui sont le résultat de l'opération de recherche.

La recherche de base de Lucene peut être effectuée à l'aide des classes suivantes :

- « IndexSearcher » est la classe donnant accès aux indexes en recherche. Elle lit ou recherche les index créés après le processus d'indexation.
- « Analyzer » : les analyseurs font parties du processus de recherche afin de normaliser les critères de recherche.
- « QueryParser » : un analyseur de requêtes. Elle est responsable de la traduction de la requête de recherche en informations que la machine peut comprendre.
- « Query » : représente la requête de l'utilisateur et elle est utilisée par IndexSearcher. Query est une classe abstraite qui contient diverses méthodes utilitaires et est le parent de tous les types de requêtes que Lucene utilise pendant le processus de recherche.

- « Hits » : est la collection d'éléments résultats de la recherche.
- « Hit » : un élément de la collection des résultats.
- « Document » : le document retrouvé et tel qu'il était lors de son ajout dans l'index (constitue des mêmes champs).

1.5 La collection TREC microblogs 2011 :

Dans le cadre de la piste de microblog TREC 2011, Twitter a fourni des identifiants pour environ 16 millions de tweets échantillonnés entre le 23 janvier et le 8 février 2011. Le corpus est conçu pour être un échantillon réutilisable et représentatif de la twitterphère - c'est-à-dire que les tweets importants et les spams sont inclus.

1.6 Trec_eval :

Le trec_eval est un outil utilisé pour évaluer les classements, que ce soit des documents ou toute autre information triée par pertinence. L'évaluation est basée sur deux fichiers : le premier, appelé « qrels » (pertinence des requêtes), répertorie les jugements de pertinence pour chaque requête. Le second contient les classements des documents renvoyés par votre système RI.

— Téléchargement :

trec_eval est téléchargeable à partir de : « **https ://trec.nist.gov/trec_eval/** »

Une fois téléchargé, extraire les fichiers dans un dossier et taper « **make** » dans l'invite de ligne de commande pour compiler le code source trec_eval.

Après cela, l'exécutable trec_eval sera prêt à être utilisé.

— Utilisation :

La commande pour exécuter trec_eval a le format suivant :

$ ./trec_eval [-q] [-m measure] qrel_file results_file

Où :

- trec_eval : est le nom du programme exécutable ;

- -q : en plus de l'évaluation récapitulative, donner une évaluation pour chaque requête ou sujet ;
- qrel_file : chemin du fichier avec la liste des documents pertinents pour chaque requête ;
- -m : affiche uniquement une mesure spécifique (« -m all_trec » montre toutes les mesures, « -m official » est le paramètre par défaut qui n'affiche que les mesures principales.) ;
- results_file : chemin du fichier avec la liste des documents récupérés par votre application.

2. Implémentation de l'approche proposée :

2.1 Les classes implémentées :

Afin de réaliser l'implémentation de notre approche, nous avons étendu Lucene avec les deux classes suivantes : La classe Fonction et la classe RechercheBoosting.

2.3.1 La classe fonction :

Dans cette classe, nous avons implémenté une fonction appelée calculerInfl; qui calcule et retourne l'influence d'un utilisateur en lui fournissant son ID en paramètre. Voici un aperçu de cette classe :

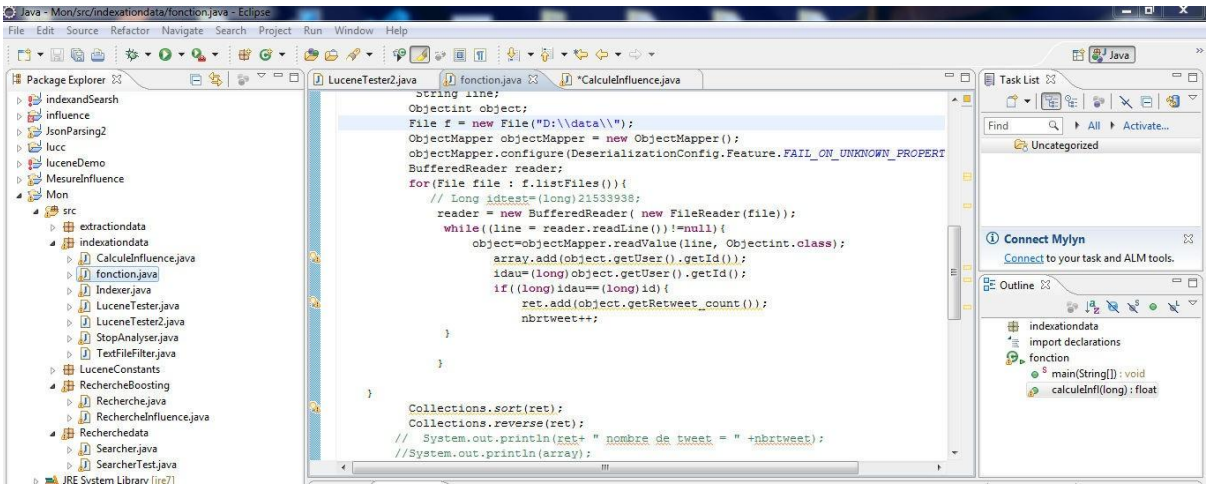


Figure 4.5 : Récupération des valeurs de retweets de chaque tweet d'un twitto

Une fois les retweets sont récupérés, nous les ajoutons dans un ArrayList pour pouvoir ensuite calculer les ratios de retweet et somme de retweet et ainsi déterminer le h_index d'un twitto. Voici la boucle qui s'en charge :

```
        Collections.sort(ret);
        Collections.reverse(ret);
        for(int i=0; i<ret.size(); i++){
            if((Integer)ret.get(0)==0){
                 break;
             }
            if(i+1<=(Integer)ret.get(i)){
                 h_index=h_index+1;
             }
            ret_count=(Integer) ret.get(i);
             if(ret_count!=0)tweetretweeté++;
             sommeret=sommeret+(Integer) ret.get(i);
        }

        ratio_retweet=(float)((float)((float)tweetretweeté+1)/(float)((float)nbrtweet+1));
        ratio_somme=(float)((float)((float)sommeret+1)/((float)(float)nbrtweet+1));
        score_total=(float)(ratio_somme*ratio_retweet*(float)h_index);

    return score_total;
}
```

Figure 4.6 : calcule du score total d'influence

Cette fonction est appelée par l'indexer pour récupérer l'influence de chacun des twitto, qui est ensuite indexée et stockée dans un field lors de l'indexation.

```
private Document getDocument( Object object) throws IOException {
     Document document = new Document();
     float infl_twitto=0;
     fonction rat=new fonction();
     infl_twitto=rat.calculeInfl((long)object.getUser().getId());

     //index file contents
     Field contentField = new Field(LuceneConstants.CONTENT, object.getText(),Field.Store.NO
     //index file name
     Field id_tweet = new Field(LuceneConstants.ID_TWEET, object.getId()+"",Field.Store.YES,
     //index auteur
     Field auth = new Field(LuceneConstants.AUTH, object.getUser().getName(),Field.Store.YES
     //index Id_auteur
     Field id_auth = new Field(LuceneConstants.ID_AUTHEUR, object.getUser().getId()+"",Field
     //index influence
     Field influence=new Field(LuceneConstants.INFLUENCE, infl_twitto+"",Field.Store.YES,Fie


     document.add(contentField);
     document.add(id_tweet);
     document.add(auth);
     document.add(id_auth);
     document.add(influence);
     return document;
```

Figure 4.7 : Récupération et indexation de l'influence

2.3.2 La classe RechercheBoosting :

Cette classe étend de la classe CustomScoreQuery de Lucene. Dans cette classe, nous ajoutons le score social (influence) au score thématique. Pour se faire, nous commençons par récupérer les fields nécessaires tels que le nom de l'auteur et son influence de l'index.

```
private class RecencyBooster extends CustomScoreProvider {
    final String[] autheur;
    final float[] influence;
    final int[] id_autheur;
    public RecencyBooster(IndexReader r) throws IOException  {
         super(r);
         autheur = FieldCache.DEFAULT.getStrings(r, LuceneConstants.AUTH);
         influence=FieldCache.DEFAULT.getFloats(r, LuceneConstants.INFLUENCE);
         id_autheur=FieldCache.DEFAULT.getInts(r, LuceneConstants.ID_AUTHEUR);
    }
```

Figure 4.8 : Fonction qui récupère les valeurs des fields

Puis à partir de ces valeurs, nous ajoutons le score social de chaque tweet à son score thématique et nous le retournons. Cette classe est appelée au moment de la recherche.

```
public float customScore(int doc, float subQueryScore,float valSrcScore) {
    float score = 0;
    String auth= autheur[doc];
    float infl_twitto=influence[doc];
    int idau=id_autheur[doc];
    if(auth!=null){
        score= (float)infl_twitto;
        return  0.1f*subQueryScore + 0.9f*score;
    }
    else return subQueryScore;
}
```

Figure 4.9 : Fonction qui retourne le nouveau score

3. Evaluation : Tests et Résultats :

Afin de réaliser nos tests, nous avons utilisé une collection de tests réduite de la collection TREC microblogs 2011 que nous détaillerons dans la séquence 3.1.1. Dans un premier temps, nous avons réalisé une recherche thématique, par la suite une recherche thématique en lui ajoutant le score sociale.

3.1 Protocole d'évaluation :

Dans cette séquence, nous présentons la collection sur laquelle se sont portés nos tests, et les mesures standard qui nous ont aidé à évaluer notre approche.

3.1.1 Notre collection de tests :

La collection réduite de la collection TREC microblogs 2011 utilisée contient :

— 2 millions de Tweets ;

— 49 requêtes ;

— Des jugements de pertinence associés à ces requêtes.

De cette collection nous avons extrait une sous collection composée de :

— 27 170 tweets ;

— 7 requêtes ;

— Les jugements de pertinences associés à ces requêtes.

3.1.2 Mesures d'évaluation utilisées :

Afin de réaliser l'évaluation de notre approche, nous avons utilisé trec_eval cité plus haut, cet outil évalue les 1000 premiers résultats retournés par le système en mettant en œuvre des mesures d'évaluation standard. Pour notre cas, nous avons utilisé les mesures d'évaluation suivantes :

— La MAP ;

— La R-précision ;

— La précision@X ;

— La courbe rappel/précision.

3.2 Résultats :

Dans cette partie, nous présentons les résultats obtenus lors des expérimentations de notre approche, en les comparant avec les résultats de l'approche thématique.

3.2.1 Résultats avec le score thématique :

Voici quelques résultats obtenus suite à la recherche thématique effectuée sur les requêtes de notre collection :

```
Q1 29589606894669824  0  3.1300497
Q1 30232528631635968  1  2.5821593
Q1 29590789294133249  2  2.0866997
Q1 29591227036860416  3  2.0866997
Q1 29956360107986944  4  2.0866997
Q1 30231301034348546  5  2.0866997
Q1 30231948798459904  6  2.0866997
Q1 30232475246526465  7  2.0866997
Q1 30230584039051264  8  2.0866997
Q1 30233817738379264  9  2.0866997
```

Figure 4.10 : Résultat du score thématique

3.2.2 Résultats obtenus en ajoutant le score social (l'influence) :

Voici quelques résultats obtenus suite à la recherche effectuée avec le nouveau score sur les requêtes de notre collection :

```
Q1    29584811014230016  0  18.182587
Q1    30230584039051264  1  3.8086698
Q1    29589606894669824  2  0.31300497
Q1    30232528631635968  3  0.25821593
Q1    29590789294133249  4  0.20866998
Q1    29591227036860416  5  0.20866998
Q1    29956360107986944  6  0.20866998
Q1    30231301034348546  7  0.20866998
Q1    30231948798459904  8  0.20866998
Q1    30232475246526465  9  0.20866998
```

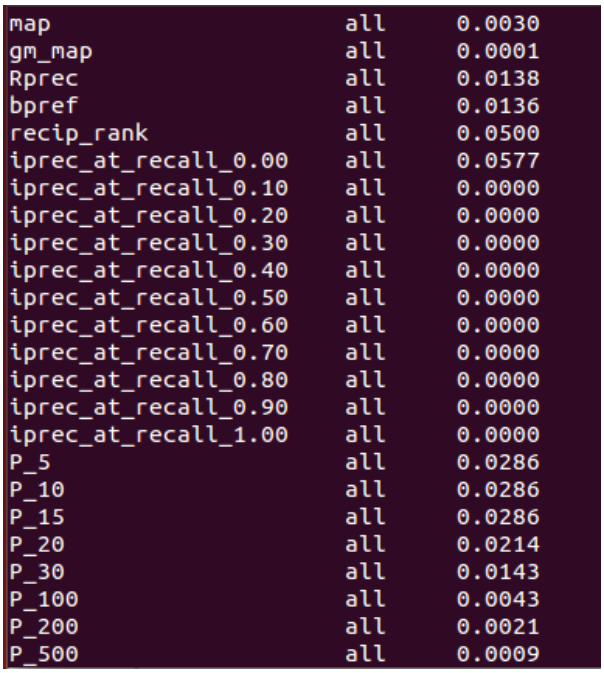
Figure 4.11 : Résultat du score d'influence

3.2.3 Évaluation des résultats :

Voici les résultats d'évaluation retournés par trec_eval :

```
map                     all     0.0010
gm_map                  all     0.0001
Rprec                   all     0.0036
bpref                   all     0.0129
recip_rank              all     0.0343
iprec_at_recall_0.00    all     0.0343
iprec_at_recall_0.10    all     0.0000
iprec_at_recall_0.20    all     0.0000
iprec_at_recall_0.30    all     0.0000
iprec_at_recall_0.40    all     0.0000
iprec_at_recall_0.50    all     0.0000
iprec_at_recall_0.60    all     0.0000
iprec_at_recall_0.70    all     0.0000
iprec_at_recall_0.80    all     0.0000
iprec_at_recall_0.90    all     0.0000
iprec_at_recall_1.00    all     0.0000
P_5                     all     0.0286
P_10                    all     0.0143
P_15                    all     0.0190
P_20                    all     0.0143
P_30                    all     0.0143
P_100                   all     0.0043
P_200                   all     0.0021
P_500                   all     0.0009
```

Figure 4.12 : Résultat d'évaluation de la recherche thématique



```
map                     all     0.0030
gm_map                  all     0.0001
Rprec                   all     0.0138
bpref                   all     0.0136
recip_rank              all     0.0500
iprec_at_recall_0.00    all     0.0577
iprec_at_recall_0.10    all     0.0000
iprec_at_recall_0.20    all     0.0000
iprec_at_recall_0.30    all     0.0000
iprec_at_recall_0.40    all     0.0000
iprec_at_recall_0.50    all     0.0000
iprec_at_recall_0.60    all     0.0000
iprec_at_recall_0.70    all     0.0000
iprec_at_recall_0.80    all     0.0000
iprec_at_recall_0.90    all     0.0000
iprec_at_recall_1.00    all     0.0000
P_5                     all     0.0286
P_10                    all     0.0286
P_15                    all     0.0286
P_20                    all     0.0214
P_30                    all     0.0143
P_100                   all     0.0043
P_200                   all     0.0021
P_500                   all     0.0009
```

Figure 4.13 : Résultat d'évaluation de notre approche

Tableau 4.1 : Les résultats de R-précision et MAP.

Score	Thématique	Thématique +influence
MAP	0.0010	0.0030
R-précision	0.0036	0.0138

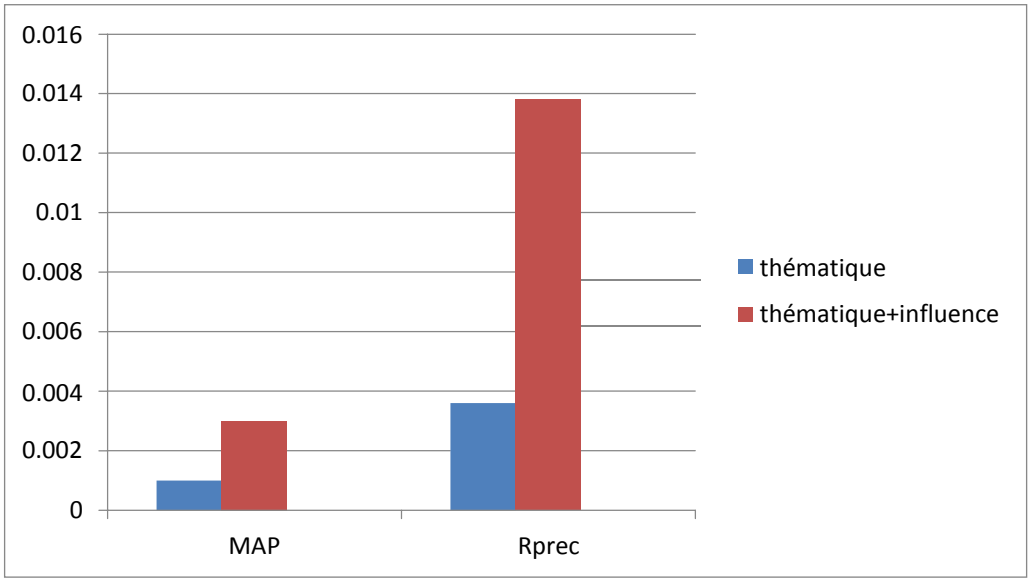


Figure 4.14 : Les résultats de R-précision et MAP.

Selon le tableau, notre approche améliore nettement les résultats par rapport à la thématique en comparant le MAP et la R-précision. La MAP de notre approche est plus importante que celle du score thématique de 0.0020 et le même constat est fait avec la R-précision où il y a une amélioration de 0.0102.

Tableau 4.2 : Les résultats de la précision@X.

score	p@5	p@10	p@15	p@20	p@30	p@100	p@200	p@500	p@1000
Thématique	0.0286	0.0143	0.0190	0.0143	0.0143	0.0043	0.0021	0.0009	0.0004
Thématique+ influence	0.0286	0.0286	0.0286	0.0214	0.0143	0.0043	0.0021	0.0009	0.0004

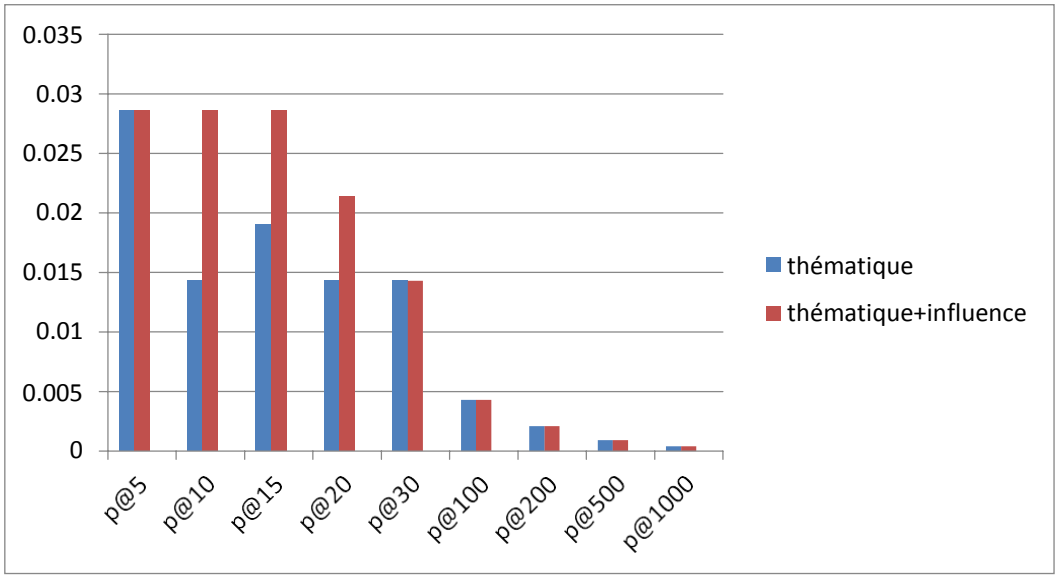


Figure 4.15 : Les résultats de la précision@X.

Notre approche a donné des résultats similaires ou meilleurs par rapport à la thématique en comparant les P@X des deux approches. Par exemple, pour les 10 premiers tweets retournés, il y a une amélioration de 0.0143.

Tableau 4.3 : Les résultats de la courbe rappel/précision.

Rappel	**0**	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**	**0.9**	**1**
Précision thématique	0.0343	0	0	0	0	0	0	0	0	0	0
Précision thématique+ influence	0.0577	0	0	0	0	0	0	0	0	0	0

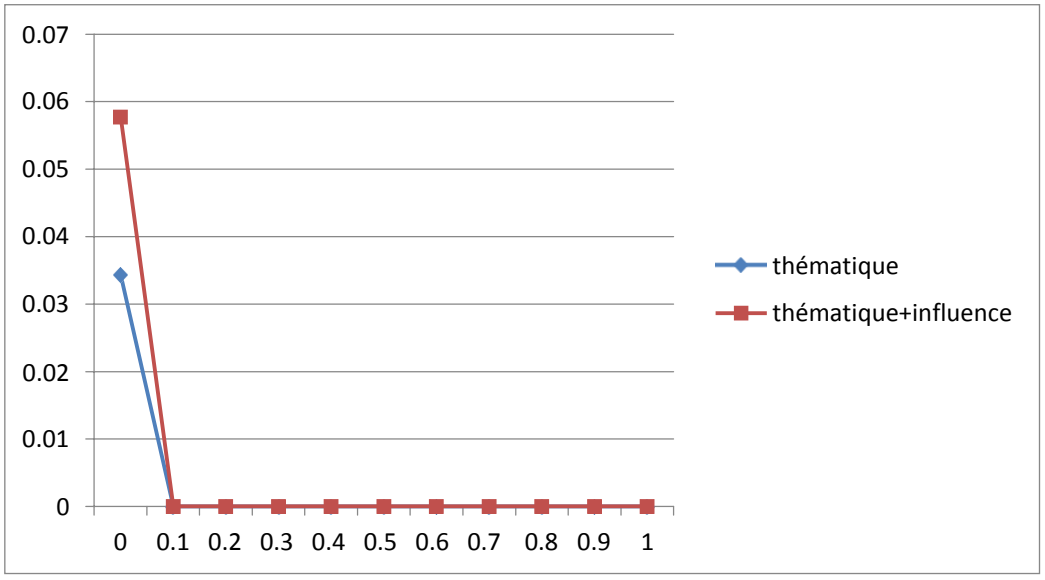


Figure 4.16 : Courbe rappel/précision.

Selon la courbe rappel/précision, notre approche améliore nettement les résultats par rapport à la thématique.

— Synthèse :

Enfin, d'après les résultats obtenus, nous concluons que notre approche qui intègre l'influence d'un twitto apporte une nette amélioration par rapport à l'approche thématique. La collection sur laquelle nous avons effectué nos expérimentations n'était pas volumineuse mais malgré cela, nous avons obtenu des résultats satisfaisants.

Conclusion :

Dans ce dernier chapitre, nous avons proposé le cadre expérimental de notre approche que nous avons présenté dans le chapitre précédent. Ensuite, nous avons réalisé des expérimentations sur un corpus de test, nous avons analysé les résultats et nous les avons comparés aux résultats thématiques. Nous concluons que notre approche, qui intègre l'influence d'un twitto, apporte une nette amélioration par rapport à l'approche thématique. Nous avons obtenu des résultats satisfaisants. Cela démontre la pertinence de notre approche et nous encourage à améliorer notre approche afin d'aboutir à de meilleurs résultats.

Conclusion générale

Conclusion générale

Dans notre travail, nous nous sommes intéressées à la recherche d'influenceurs dans le réseau social Twitter. L'objectif étant de retrouver les personnes les plus influentes sur ce dernier. Pour cela, nous avons proposé une approche dans le cadre de la recherche d'influenceurs dans Twitter où nous avons exploité la relation de retweet ainsi qu'une mesure bibliométrique : le h-index que nous avons adapté à Twitter. Par la suite nous avons implémenté cette approche, montré quelques résultats de tests liés à la recherche thématique et évalué ces résultats selon différentes métriques. Enfin, nous avons essayé de démontrer l'efficacité de cette approche dans la recherche d'information sur Twitter.

A l'avenir, nous envisagerons de tester notre approche sur une collection encore plus volumineuse.

Ce projet nous a permis d'enrichir nos connaissances théoriques et personnelles et développer nos propres capacités notamment sur le plan pratique.

Bibliographie:

[Anger. 11] Kittl C. Anger I. « Measuring influence on twitter. In S. N. Lindstaedt, M. Granitzer (Eds.), ». 2011.

[Azaza et al. 15] Savonnet M. Frame A. Azaza L. Kirgizov S. « Évaluation de l'influence sur Twitter : Application au projet « Twitter aux Elections Européennes 2014 ». Mai 2015.

[Badache. 16] Ismail Badache. « Recherche d'information sociale : exploitation des signaux sociaux pour améliorer la recherche d'information, ». 2016.

[Benjabeur L. 11] Benjabeur L., Tamine L., Boughanem M. «Un modèle de recherche d'information sociale dans les microblogs : cas de Twitter». 2011

[Bernard Bathlot, 17] Bernard Bathlot. «Leader d'opinion». 12 Octobre

2017. https://www.definitions-marketing.com/definition/leader-d-opinion/

[Camille Jourdain. 09] Camille Jourdain. «Médias sociaux – Comment mesurer l'influence sur Twitter ?»**.** 5 décembre 2009.

https://www.camille jourdain.fr/medias-sociaux-comment-mesurer-linfluence-sur-twitter/+&cd=1&hl=fr&ct=clnk&gl=dz

[Cha. 10] Haddadi H. Benevenuto F. Gummadi P. K. Cha M. Measuring user influence in twitter: The million-follower fallacy. In. 2010.

[Damak. 14] Firas Damak. «Étude des facteurs de pertinence dans la recherche de microblogs». 2014.

[Ding. 13] Jia Y. Zhou B. Han Y. He L. Zhang J. Ding Z. Measuring the spreadability of users in microblogs. 2013.

[F. Boubekeur. 17] M. Ferrouk et L. Belkacem F. Boubekeur. « Nouvelle mesure de l'influence sur twitter. ». 2017.

[Fabian Riquelem. 16] Pablo Ganzalez-Cantergiani Fabian Riquelem. « Measuring user influence on Twitter : A survery ». 2016.

[Hammache. 13] Arezki Hammache. «Recherche d'Information : un modèle de langue combinant mots simples et mots composés,». 2013.

[Haewoon Kwak. 10] Hosung Park Haewoon Kwak Changhyun Lee & Sue Moon. «What is Twitter, a Social Network or a News Media? ». 2010.

[J. Rocchio. 71] Rocchio J.J. Relevance Feedback in Information Retrieval, In The SMART System Experiments in Automatic Document Processing, 1971.

[K. Mechach, 16] Mechach Kheira. « Etude de l'impact des methodes de localisation dans les systèmes d'information distribués, ». 2016.

[Kleinberg. 99] Kleinberg J. Authoritative sources in a hyperlinked environment. Journal of the ACM (JACM), vol. 46, n° 5, p. 604-632. 1999.

[Manning et al., 08] Manning, C. D., Raghavan, P., and Schutze, H. Introduction to Information Retrieval. Cambridge University Press; 1st edition. 2008.

[Nagmoti et al., 10] Nagmoti, R., Teredesai, A., et De Cock, M. Ranking approaches for microblog search. In Proceedings of the 2010 ieee/wic/acm international conference on web intelligence and intelligent agent technology (pp. 153‑157). Washington, USA : IEEE Computer Society. 2010.

[Romero D. M. 11] Asur S. Huberman B. A. Romero D. M. Galuba W. « Influence and Passivity in Social Media ». 2011.

[Salton et al., 88] Salton, G., and Buckley, C. Term-weighting approaches in automatic text retrieval. Information Processing & Management (IPM) 24, 5 (1988), 513–523.

[Salton, 88] Salton, G. Syntactic approaches to automatic book indexing. In Proc. Of the annual meeting on Association for Computational Linguistics (ACL) (1988), Department of Computer Science, Cornell University, Ithaca, New York, pp. 204-2010.

[Shamma et al.. 09] Shamma, D. A., Kennedy, L., et Churchill, E.F. Tweet the debates: Understanding community annotation of uncollected sources. In proceedings of the first sigmm wordkshop on social media (pp. 3-10). New York, NY, USA: ACM. 2009.

[Sung J. 13] Moon S. Sung J. & Lee JG. « The influence in Twitter: Are they really influenced? ». 2013.

[Tunklang. 09] Daniel Tunklang « A Twitter Analog to PageRank ». 2009.

http://thenoisychannel.com/2009/01/13/a-twitter-analog-to-PageRank.

[Weng J. 10] Jiang J. He Q. Weng J. Lim E.-P. « TwitterRank: finding topic-sensitive influential Twitterers ». 2010.

[Zhang M. 11] Zhang M., Sun C., Liu W. Identifying Influential Users Of Micro Blogging Services : A Dynamic Action-Based Network Approach. PACIS 2011; Proceedings.223; 2011.

[Zhao et al., 11] Zhao, L., Zeng, Y., et Zhong, N. 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. 2011.

Annexes

H-index :

Définition :

Nommé d'après son créateur Jorge H. Hirsch, le h-index évalue la production scientifique des chercheurs sur base des citations au sein d'un corpus donné. Un h-index de N signifie que le chercheur considéré (ou le groupe de chercheurs) a publié N articles cités au moins N fois dans la base de données considérées.

Exemple :

Un h-index de 6 signifie que 6 publications de l'auteur ont chacune été cités au moins 6 fois.

Calcul de h-index :

Le h-index est calculé en classant et en numérotant les publications de l'auteur de la plus citée (n°1) à la moins citée. Le h-index correspond au dernier numéro de la publication qui vérifie :

$$numéro\ de\ la\ publication \leq nombre\ de\ \ citations.$$

Dans l'exemple ci-dessous (tableau 1), l'indice H est de 6, car six articles ont reçu au moins six citations.

Tableau A.1 : Exemple de h-index

Articles	Nombre de Citations	
1er	77	
2e	35	
3e	33	
4e	10	
5e	8	
6e	7	h-index = 6
7e	2	

Le h-index d'un chercheur est de plus en plus demandé dans les dossiers de soumission aux appels à projet ou dans le cadre d'évaluations de la recherche.