**R**épublique **A**lgérienne **D**émocratique et **P**opulaire
**M**inistère de l'**E**nseignement **S**upérieur et de la **R**echerche **S**cientifique

**U**NIVERSITE **M**OULOUD **M**AMMERI DE **T**IZI-**O**UZOU

**F**ACULTE DU **G**ENIE **E**LECTRIQUE ET D'INFORMATIQUE
**DEPARTEMENT D'INFORMATIQUE**

# Mémoire de Fin d'Etudes de MASTER ACADEMIQUE

Domaine : **Mathématiques et Informatique**
Filière : **Informatique**
Spécialité : **Conduite de projets informatiques**

*Présenté par*
**Ahcene DJEDDI**
**Amir BENDOU**

**Thème**

# Recherche d'information temporelle dans les microblogs cas : Twitter

*Mémoire soutenu publiquement le 10/07/ 2016 devant le jury composé de :*

**Président : M$^{me}$ Fatiha AMIROUCHE**
**Encadreur : M$^{me}$ Lila BELKACEMI**
**Examinateur : M$^{elle}$ Wassila AZZOUG**
**Examinateur : M$^{r}$ Mohamed Nabil AMIROUCHE**

# Dédicace

*Je dédie ce modeste travail*

*A mes parents en guise de reconnaissance et de gratitude pour les sacrifices qu'ils ont faits.*

*A mon frère, à ma sœur, à qui je dois tout l'amour, avec tous mes vœux de les voir réussir dans leurs vies.*

*A toutes la famille Djeddi et la famille Tabti.*

*A mes ami(e)s, à qui je souhaite le succès pour l'amitié qui nous a toujours unis.*

*A tous ceux qui me sont chers.*

DJEDDI Ahcene

# Dédicace

*Je dédie ce modeste travail*

*A mes parents en guise de reconnaissance et de gratitude pour les sacrifices qu'ils ont faits.*

*A mon frère à ma sœur, à qui je dois tout l'amour, avec tous mes vœux de les voir réussir dans leurs vies.*

*A toutes la famille Bendou et la famille Djouahra.*

*A mes ami(e)s, à qui je souhaite le succès pour l'amitié qui nous a toujours unis.*

*A tous ceux qui me sont chers.*

*BENDOU Amir*

# Sommaire

**Chapitre IV Evaluation expérimentale**

# Liste des figures :



# Liste des tableaux :

# Introduction Générale

La croissance d'Internet a permis de former différents types de réseaux sociaux(RS) à grande échelle qui sont maintenant reconnus comme un moyen important pour la diffusion de l'information. Avec les RS les internautes ne sont pas que des consommateurs d'information, mais aussi des producteurs. Ils consultent, créent, partagent et diffusent de l'information.

De la famille des RS, nous avons les plateformes de microblogging. Les plateformes de microblogging permettent aux microbloggeurs de publier des informations sur différents sujets : des opinions, des événements, des statuts. . . Parmi les plate-formes de microblogging, Twitter est sans conteste la plate-forme la plus utilisée (Damak,2014).

Ce nouveau contexte de diffusion de l'information, c'est-à-dire les réseaux sociaux, sur le Web peut constituer un moyen efficace pour cerner les besoins en information des utilisateurs du Web. C'est en tous les cas ce que la recherche d'information sociale tente de prouver. La recherche d'information sociale(RIS) est donc un nouveau paradigme de recherche. Elle consiste à adapter les modèles et les algorithmes de la RI classique en exploitant les sociales issues des réseaux sociaux.

Nous nous intéressons dans ce travail à la recherche d'information dans les microblogs. Les modèles de RI classiques, conçus pour des textes plus longs que les 140 caractères d'un microblog, ne sont plus adaptés pour ces derniers.

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

# Chapitre I :
# La recherche d’information classique

# Introduction

La recherche d'information(RI) n'est pas un domaine récent, il date des années 1940, dès la naissance des ordinateurs. A ses débuts, la RI était liée aux applications dans les bibliothèques. Et avec l'avènement du web et la quantité d'information qui ne cesse d'augmenter, ce domaine n'est pas uniquement réservé aux bibliothèques mais à tout le web.

La recherche d'information (information retrieval en anglais) est un domaine qui consiste à définir des modèles et des processus dont le but est de retourner, à partir d'un corpus de documents indexés, ceux dont le contenu correspond le mieux au besoin en information exprimé par un utilisateur.

# I. Système de recherche d'information(SRI)

## I.1 Définition

Un système de recherche d'information(SRI) est un programme informatique qui permet de sélectionner, dans une collection de documents, ceux qui sont Susceptibles de répondre à un besoin en information exprimé par une requête utilisateur.

## I.2 Processus de recherche d'information

La recherche d'information implique le stockage, la recherche et l'exploration des documents pertinents. Pour cela, plusieurs concepts sont utilisés :

- **Documents** : le document représente l'unité élémentaire d'information accessible et exploitable par le SRI et peut constituer une réponse aux besoins utilisateur. Le document peut être un texte, une page WEB, une carte, un article sur un blog ou un micro-blog, une image, une vidéo.

- **Collection de documents :** c'est un ensemble de documents réunit dans un corpus, facilement accessibles et exploitables pour un souci d'optimalité. La

base constitue des représentations simples pour ces documents, et étudiées de telles sortes que la recherche et la gestion se font dans les meilleures conditions de coût.

- **Requête :** la requête est la représentation du besoin en information de l'utilisateur exprimé sous un langage particulier.

Dans le but de trouver les documents pertinents qui répondent au mieux à la requête le SRI utilise un processus qui est constitué de deux phases principales : l'indexation et l'appariement requête/document.

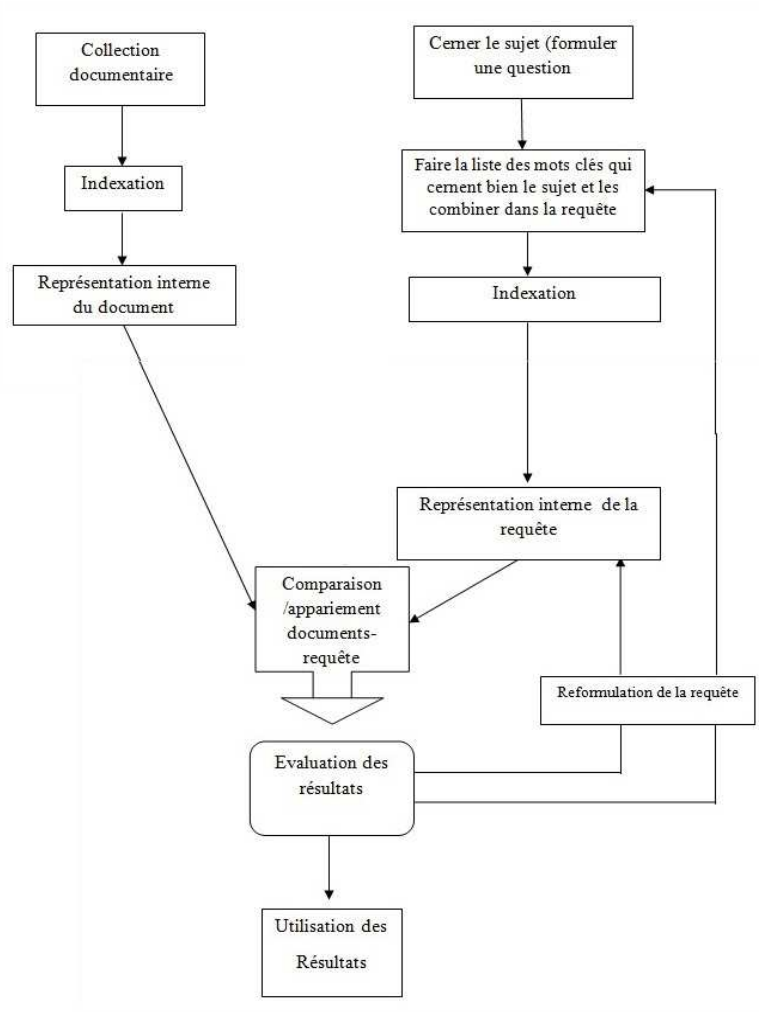


**Figure 1-1 : Processus général de recherche d'information**

## I.2.1 L'indexation

Les documents dans leurs états bruts sont coûteux et difficile à exploiter c'est pour remédier à ce problème que les SRI utilise l'indexation.

L'indexation consiste à constituer le descriptif (mots-clés) qui représente le mieux le document ou la requête assorti d'un poids il est ensuite stocké dans une structure appelé index qui peut être facilement interrogée.

Cette opération peut-être assez longue en fonction du nombre de documents de la collection, de leurs tailles et du type indexation (manuelle, semi-automatique ou automatique.).

a. **Indexation manuelle:** l'indexation manuelle est réalisée par un expert ou un documentaliste qui lit chaque document et fournis une terminologie spécifique pour indexer chaque document.

   Cette manière d'indexer garantie une meilleure représentation d'un document. Cependant elle est fastidieuse et coûteuse en temps pour une grande collection, et le facteur humain fait entrer un degré du subjectivité qui fait qu'un même document peut être indexé de façon différentes par des personnes différentes par la même personne à des moments différents.

b. **Indexation automatique*:*** *le processus d'indexation est entièrement automatisé.* La majorité des SRI suive cette indexation en raison de sa rapidité et des gains en coûts par rapport à l'indexation manuelle.

c. **Indexation semi-automatique :** cette approche est une combinaison de l'indexation manuelle et automatique. Elle se base sur l'indexation automatique puis une intervention humaine est réalisée pour faire le choix sur les termes significatifs et valider la représentation finale.

L'indexation automatique est composée d'une chaîne de traitement qu'on applique sur chaque document et aussi sur les requêtes.

1. **L'analyse lexicale:** l'analyse lexicale constitue la première étape du processus d'indexation. Lors de cette étape un document textuel est transformé en un ensemble de termes. La ponctuation, la casse, et la mise en

page sont supprimées.

2. **Élimination des mots vides:** cette opération supprime les termes non significatifs (mots vides) tels que les pronoms personnels, les articles, les mots de liaison, ou les prépositions. Les mots dépassants un certain nombre d'occurrences dans le document sont aussi supprimés.
3. **Lemmatisation (radicalisation*):*** *cette phase* permet de substituer chaque mot par sa racine. Parmi les techniques utilisées dans la lemmatisation, nous citons :
    - ✓ La table de consultation (dictionnaire).
    - ✓ L'élimination des affixes (Porter).
    - ✓ La troncature.
    - ✓ Les variétés de successeurs (n grammes).
    - ✓ L'utilisation des étiqueteurs grammaticaux (taggeurs).

4. Pondération : une fois les termes identifiés et normalisés, vient l'étape de pondération qui consiste à affecter un poids pour chaque terme. Ce poids est une valeur numérique qui représente l'importance du terme dans le document. La plupart de ces calcule utilise les facteurs TF et IDF, qui combine les pondérations locales (dans le document) et globales (dans la collection). La pondération d'un terme s'exprime en fonction de deux pondérations :

    - ***Tf (Term Frequency)*:** l'idée est que plus un terme est fréquent dans un document plus il est important dans la description de ce document.
    - ***Idf (Inverse of Document Frequency)*:** cette mesure calcule la fréquence d'un terme dans la collection (pondération globale) dans le but d'identifier les termes qui discriminent le plus un document par rapport aux autres documents de la collection.

## I.2.2 L'appariement

Cette étape vient après l'indexation des documents et l'analyse de la requête. Le SRI prédit les documents que l'utilisateur trouvera pertinent par la mise en correspondance de la requête et de l'index puis calcule un score de pertinence qui reflète le degré de similarité entre la requête et le document. Ce score est calculé à partir d'une valeur ***RSV (q, d)*** (Retrieval Status Value), où ***q*** est une requête et ***d*** un document. Cette mesure tient compte de la pondération du terme.

Afin d'améliorer la capacité d'un système de recherche d'information à restituer les documents pertinents pour l'utilisateur, un troisième processus peut être intégré : ***processus de reformulation de requête***.

## I.2.3 Reformulation de requête

La reformulation consiste à réajuster les poids des termes de la requête ou à rajouter des termes reliés à ceux de la requête initiale. Elle peut être manuelle (avec intervention de l'utilisateur) ou automatique.

La stratégie de reformulation de la requête la plus connue est la reformulation par ***réinjection de pertinence (Relevance Feedback***). Le principe général de cette stratégie se résume ainsi :

**a)** L'utilisateur formule sa requête (requête initiale) ;

**b)** Présenter, à l'utilisateur, les documents retournés par le SRI en réponse à la requête initiale;

**c)** L'utilisateur sélectionne à partir de cette liste les documents qu'ils lui conviennent;

**d)** La requête de départ est alors modifiée pour tenir en compte des jugements de l'utilisateur

# II. Modèles de recherche d'information

## II.1 Le modèle booléen

Ce modèle est basé sur la théorie des ensembles. Dans ce modèle chaque document ***d*** est représenté comme une conjonction logique de termes (non pondérés) ***d = t1, t2, ...tn.*** Une requête ***q*** est une expression logique dont les termes d'indexation sont assemblés par les opérateurs logiques: ∧**(conjonction),** ∨**(disjonction)** et ¬**(négation)**, par exemple

***q = (t1∧ t2) ∨ ¬t3.***

La fonction d'appariement *RSV* est la vérification de l'implication logique ***d→q***. Ainsi, les documents qui satisfont l'expression logique qui représente la requête sont considérés comme pertinents.

Le modèle booléen utilise le mode d'appariement exact qui consiste à ne restituer que les documents répondant exactement à la requête.

***RSV (document, requête) = {1,0}.***

Ce modèle est très simple et aussi très utilisé mais possède deux inconvénients :

- ✓ Les requêtes sont difficile a représenter sous forme d'expression pour les utilisateurs, et
- ✓ L'appariement des documents est pondéré avec un poids (0ou1) donc tous les documents on la même pertinence et on ne peut pas les classer selon un ordre du plus proche du besoin.

## II.2 Le modèle vectoriel

Le modèle vectoriel est ce modèle algébrique qui représente les documents et les requêtes sous forme de vecteurs de poids dans l'espace vectoriel des termes d'index.

Dans un espace d'information à m termes T={$t_1$,$t_2$,….., $t_m$}, un document $\boldsymbol{d}_i$ est représenté par un vecteur de poids $\boldsymbol{w_{ij}}$ de dimension **m** dans l'espace vectoriel composé de tous les termes d'indexation $\boldsymbol{d_i = (w_{i1}, w_{i2}, ... ,w_{im})}$. Une requête ***q*** est aussi représentée par un vecteur de poids $\boldsymbol{w_q}$ défini dans le même espace vectoriel que le document

$\boldsymbol{q=(w_{q1}, w_{q2}, \ldots, w_{qm})}$ où $\boldsymbol{w_{qj}}$ est le poids de terme $\boldsymbol{t_j}$ dans la requête $\boldsymbol{q}$, et $w_{ij}$ son poids dans le document $\boldsymbol{d_i}$. Le modèle vectoriel prend en compte le poids de terme dans le document .Ce dernier peut être soit :

- ✓ une forme de tf*idf,
- ✓ un poids attribué manuellement par l'utilisateur.

La pertinence du document $\boldsymbol{d_i}$ pour la requête $\boldsymbol{q}$ est mesurée par le degré de corrélation de leurs vecteurs correspondants. Cette corrélation peut être exprimée par l'une des mesures suivantes :

| Mesures | Formules |
|---|---|
| Le produit scalaire | $RSV(q,d_i)=\sum_{j=1}^{\lvert T\rvert} w_{qj}.w_{ij}$ |
| La mesure de cosinus | $RSV(q,d_i)=\frac{q.d_i}{\lVert q\rVert.\lVert d_i\rVert}=\frac{\sum_{j=1}^{\lvert T\rvert} w_{qj}.w_{ij}}{\sqrt{\sum_{j=1}^{\lvert T\rvert} {w_{qj}}^2 \sum_{j=1}^{\lvert T\rvert} {w_{ij}}^2}}$ |
| La mesure de Dice | $RSV(q,d_i)=\frac{2\times\sum_{j=1}^{\lvert T\rvert} w_{qj}.w_{ij}}{\sum_{j=1}^{\lvert T\rvert} {w_{qj}}^2+\sum_{j=1}^{\lvert T\rvert} {w_{ij}}^2}$ |
| La mesure de Jaccard | $RSV(q,d_i)=\frac{\sum_{j=1}^{\lvert T\rvert} w_{qj}.w_{ij}}{\sum_{j=1}^{\lvert T\rvert} {w_{qj}}^2+\sum_{j=1}^{\lvert T\rvert} {w_{ij}}^2-\sum_{j=1}^{\lvert T\rvert} w_{qj}.w_{ij}}$ |

**Tableau 1.1 : Les mesures de similarité [Hammache, 2013]**

Le modèle vectoriel de base est l'un des modèles les plus utilisés en RI. Son avantage, par apport au modèle booléen, réside dans sa capacité d'ordonner les résultats de la recherche selon leur degré de pertinence pour une requête d'utilisateur. Cependant, ce modèle suppose que les termes d'index sont indépendants, et il ne tient pas compte des relations sémantiques qui peuvent exister entre ces termes dans le même document ou la même requête **[Azzoug, 2013].**

## II.3 Le modèle probabiliste

Ce modèle est fondé sur le calcul de la probabilité de pertinence d'un document pour une requête. Le principe de base consiste à retrouver des documents qui ont en même temps une forte probabilité d'être pertinents, et une faible probabilité d'être non pertinents. Ce modèle est proposé par Maron et Kuhns (Maron et al. 60) au début des années 1960. Le score d'appariement RSV (D, Q) entre le document D et la requête Q est calculé par la formule (Robertson et al. 1994) :

$$RSV(D, Q) = P(R|D)/P(NR|D)$$

Le modèle de base a démontré son efficacité en recherche d'information, mais sont inconvénient est de considérer les termes d'indexation comme étant indépendants les uns des autres, ce qui n'est pas toujours le cas.

- **Modèle de langue :** ce modèle procède de la manière suivant:
  - ✓ Construire pour chaque document un modèle de langue Md
  - ✓ Calculer la probabilité (q/Md) qu'une requête q puis être généré le modèle de langue
  - ✓ Cette probabilité est considérée comme le score de pertinence du document pour la requête.

$$Md\ RSV\ (Q, D) = P\ (q\ /\ M\ D) = P(t\ 1, t\ 2,..., t\ n\ /\ D)$$

Mais dans cette formule un terme qui n'a apparaît pas dans le requête auras un score nul ce qui as pour effet que la probabilité du document seras nulle.

Pour remédier à cela, des techniques de lissage, dont le lissage de Laplace, le lissage de Good-Turing, le lissage Backoff, le lissage par interpolation sont utilisés. Leur principe consiste à assigner des valeurs non nulles aux termes qui n'apparaissent pas dans un document.

- **Modèle connexionniste:** le modèle connexionniste est une application de la théorie des réseaux de neurones Dans la recherche d'information. Les réseaux

de neurones sont un certain nombre de modèles qui essaient de reproduire certaines structures de base du cerveau humain dont l'objectif d'imité certaines de ses fonction

L'idée de base est que la RI est un processus associatif qui peut être représenté par propagation de signaux de la couche d'entrée vers la couche de sortie et aussi représenté les différentes relations et associations qui existent entre les termes, les documents

# III. Évaluation d'un SRI

L'évaluation d'un SRI est une étape très importante pour sa validation. Elle permet de définir les caractéristiques du système en termes de qualité de service et de facilité d'utilisation selon les critères suivants:

- ✓ le temps de réponse,
- ✓ la présentation des résultats,
- ✓ l'effort fourni par l'utilisateur pour retrouver parmi les documents retournés ceux qui sont pertinents,
- ✓ le taux de rappel du système
- ✓ le taux de la précision du système.

Le temps de réponse, la représentation des résultats et l'effort fourni par l'utilisateur sont des mesures de la qualité de service rendu à l'utilisateur tandis que le rappel et la précision essentiels aux modèles de recherche couvrent quant à eux la performance du système.

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.

# III.1 Collection de référence

Une collection de test (ou collection de référence) comprend un ensemble de documents à indexer sur lequel le système sera évalué, une liste de requêtes prédéfinies et des jugements de pertinence manuellement établis par des assesseurs humains pour chaque requête.

De nombreuses collections de référence existent en RI. Elles sont principalement mises en œuvre dans le cadre de campagnes d'évaluation, dont les principales sont :

- **La campagne CLEF :** lancée en 2000, cette campagne a pour objectif de promouvoir la recherche et le développement dans le domaine de la recherche d'information multilingue.
- **La campagne Amaryllis :** est une version française de TREC de 1996 à 1999 et sous taches de CLEF en 2002
- **La campagne INEX :** Lancée en 2002, son objectif principal est de promouvoir l'évaluation de la recherche dans les documents semi-structurés en fournissant de grandes collections de test de type XML.
- **La campagne TREC** (Text Retrieval Conférence) : constitue le projet le plus ambitieux d'évaluation des SRI. La campagne TREC est une campagne d'évaluation annuelle depuis 1992. Elle vise à explorer de nouveaux domaines de recherche et de démontrer la robustesse des méthodologies de recherche existantes. Chaque année, la campagne TREC lance de nouvelles tâches (ou pistes) correspondant aux centres d'intérêts actualisés des chercheurs de RI. Parmi ces tâches, on distingue :
    1. La tâche *spoken document retrieval*
    2. La tâche *question answering*
    3. La tâche Interactive
    4. **La tâche *Ad Hoc* :** c'est la tâche classique de la RI. Elle vise à évaluer les performances d'un SRI sur des ensembles statiques de documents.

## III.2 Les mesures d'évaluation d'un SRI

Pour évaluer la performance d'un SRI, deux principales mesures sont utilisées

- ✓ **La précision :** détermine l'aptitude d'un SRI à rejeter les documents non pertinents vis à vis à d'une requête utilisateur.
- ✓ **Le rappel :** exprime la capacité d'un SRI à sélectionner tous les documents pertinents vis à vis de cette requête.

$$\text{Précision} = \frac{\#\text{documents pertinents retrouvés}}{\#\text{documents retrouvés}}$$

$$\text{Rappel} = \frac{\#\text{documents pertinents retrouvés}}{\#\text{documents pertinents dans la base}}$$

D'autres mesures sont aussi utilisées :

✓ **La courbe précision-rappel : l**es mesures de précision-rappel ne sont pas indépendantes, en effet en réponse à une requête on a un taux de rappel égale à 1, mais une précision faible, voir de même, si on augmente la précision en restreignant le nombre de documents retournés, dans ce cas le rappel pouvant diminuer. Dans les SRI on cherche à améliorer le couple rappel et précision. Ces deux métriques ne sont pas statiques non plus (c'est-à-dire qu'un système n'a parce qu'une mesure de précision et de rappel). Le comportement d'un système varie en fonction de précision et de rappel donc de la liste ou du rang du document dans la liste. Ainsi, la courbe de la Figure suivante montre la forme générale que peut prendre la variation de rappel précision pour un système.

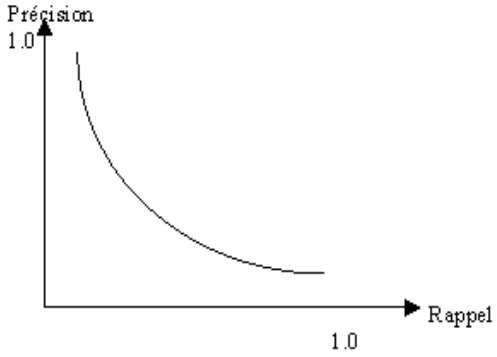


**Figure 1.2 : Courbe de précision-rappel**

La précision et le rappel croient en même temps si un document pertinent est récupéré. Plus cette courbe décroît tardivement, meilleur est l'algorithme de l'ordonnancement étudié. Cette courbe est intéressante pour comparer les résultats d'une même requête rendus par deux SRI différents. Un système dont la courbe dépasse (c'est-à-dire qu'elle se situe en haut à droite de) celle d'un autre est considéré comme un meilleur système. Il arrive parfois que les deux courbes se croisent. Dans ce cas, il est difficile de dire quel système est le meilleur.

- ✓ **F-mesure :** la F-mesure correspond à un compromis de la précision et du rappel donnant la performance du système. Ce compromis est donné de manière simple par la moyenne harmonique de la précision et du rappel de la formule suivante :

$$F_1 = \frac{2 * P * R}{P + R}$$

Où $P$ : Précision et, $R$ : Rappel.

Les mesures de rappel, précision et F-score sont des mesures basées-ensembles. Elles permettent d'évaluer des ensembles non ordonnés de résultats. On parle alors de mesures d'évaluation non ordonnées.

D'autres mesures d'évaluation ordonnées existent :

- ✓ **La précision Moyenne :** l'idée est de générer une valeur unique de ranking en moyennant les valeurs de précision obtenues après chaque document pertinent observé.
- ✓ **MAP : l**a MAP est la moyenne des précisions moyennes (Pmoy) obtenues sur l'ensemble des requêtes à chaque fois qu'un document pertinent est retrouvé :

$$MAP = \frac{\sum_{q \in Q} P_{moy}(a)}{|Q|}$$

- ✓ **R-précision : l**a R-précision est un bon paramètre pour observer le comportement d'un système pour chaque requête individuellement.

La R-précision moyenne calculée sur toutes les requêtes n'a pas d'intérêt. L'idée est de générer une valeur de ranking unique en calculant la précision au rang R, où R est le nombre de documents pertinents pour la requête courante.

$$R - Prec = P@R = \frac{|DPR|}{R}$$

## III.3 Protocole d'évaluation TREC

Pratiquement, pour évaluer un SRI, les participants à la campagne TREC doivent suivre le protocole suivant : Pour chaque requête de la collection de test fourni, les 1000 premiers documents restitués par le système sont examinées et les précisions à x points (notées P@ x), sont calculées à différents points (à 5, 10, 15, 30, 100 et 1000 premiers documents restitués).

La précision exacte découle de ces précisions. Puis, une précision moyenne MAP est calculée pour chaque requête. Il s'agit de la moyenne des précisions de chaque document pertinent pour cette requête. La précision d'un document est la précision à x, tel que x est le rang de ce document dans l'ensemble des documents pertinents retrouvés.

Finalement, les précisions moyennes pour l'ensemble des requêtes sont calculées permettant d'obtenir une mesure de la performance globale du système.

# Conclusion

Dans ce chapitre nous avons présenté les notions de base de la RI telle que les modèles de recherche et le processus de RI (indexation, appariement) et nous avons fini par présenter les étapes d'évaluation d'un SRI et les mesures utilisées pour évaluer les modèles et les systèmes de recherche.

La recherche d'information classique se base sur un processus simple qui permet de retourner les documents qu'il faut selon les besoins qui sont exprimés sous forme de requêtes et qui est très efficace pour ce qui est des documents classiques mais la RI classique deviens vite obsolète dans un contexte de microblogging.

La particularité des microblogs et l'apparition des réseaux sociaux ont mis en défi la RI. Et pour répondre aux spécificités des réseaux sociaux un nouveau type d'approches fait son apparition et ces approches sont recensées dans ce qu'on appelle la recherche d'informations sociale et c'est là l'objet de notre prochain chapitre.

# Chapitre II :

# La recherche d’information sociale

# Introduction

Les réseaux sociaux ont complètement bouleversé le web et la manière de créer du contenu, les utilisateurs sont alors passés de consommateurs passifs aux producteurs de contenus.

La recherche d'information sociale a pour objectif de retrouver des documents qui correspondent à un besoin d'information d'un utilisateur, tout en intégrant des éléments provenant de la participation des utilisateurs à des réseaux sociaux. Ces réseaux sociaux tel que Twitter offrent aux utilisateurs la possibilité de communiquer, d'interagir, et de répondre aux messages des autres.

# I. Les réseaux sociaux

Le concept de « réseau social » a été inventé en 1954 par l'anthropologue John A. Barnes. Le principe de réseau se définit par deux éléments : les contacts et les liaisons entre les contacts.

Le terme de « réseaux sociaux » (social networking) désigne l'ensemble des sites Internet permettant de constituer un réseau d'amis, de passionnés ou de connaissances professionnelles.

Des communautés d'utilisateurs se regroupent ainsi en fonction de centres d'intérêt commun. La plupart des sites qui servent de supports à ces réseaux sociaux proposent un certain nombre de fonctionnalités permettant échanges et réactivité entre membres inscrits.

Les réseaux sociaux représentent un changement de fond dans la façon dont les individus appréhendent leur vie sociale, privée ou professionnelle. En 2010, Facebook devient le site le plus visité au monde en dépassant Google. Cet événement majeur marque le début de la réussite des réseaux sociaux. Facebook (2004), Twitter (2006), LinkedIn (2003), Viadeo (2004), Skype (2003), MSN (1999) sont désormais partie intégrante de notre vie sociale.

De nos jours Il existe plus de 200 sites de réseaux sociaux, et le nombre total

d'utilisateurs actifs des réseaux sociaux dans le monde s'élève désormais à près de 2 milliards.

- ***Les chiffres clés des réseaux sociaux en 2016***
  - ✓ Sur 7,357 milliards de personnes dans le monde, on dénombre 3,715 milliards d'internautes.
  - ✓ Sur 3,715 milliards d'internautes, 2,206 milliards utilisent les réseaux sociaux chaque mois.
  - ✓ Sur 2,206 milliards d'utilisateurs des réseaux sociaux, 1,925 milliards sont actifs sur mobile.

Nous présentons dans le tableau ci-dessous, le nombre d'utilisateurs actifs de chacun des principaux réseaux sociaux :

| Réseaux social | Nombre d'utilisateurs en 2016 |
|---|---|
| Facebook | 1,5 milliard |
| YouTube | 1 milliard |
| Twitter | 500 millions |
| Google+ | 500 millions |
| LinkedIn | 400 millions |
| Instagram | 400 millions |
| Tumblr | 230 millions |
| Snapchat | 200 millions |

**Tableau 2.1 : Les réseaux sociaux les plus populaires**

# I.1 Les différents types de réseaux sociaux

- **Wiki** : un **wiki** est une application web qui permet la création, la modification et l'illustration collaboratives de pages à l'intérieur d'un site web. Il utilise un langage de balisage et son contenu est modifiable au moyen d'un navigateur web. C'est un outil de gestion de contenu, dont la structure implicite est minimale, tandis que la structure explicite émerge en fonction des besoins des usagers.
- **Blog** : le **blog** est un type de site web consacré à un sujet particulier ou une chronique personnelle où n'importe quel internaute peut donner son avis. Il

s'agit d'un espace individuel d'expression, crée pour donner la parole a tous les internautes (particuliers, entreprises, artistes, hommes politiques, associations. . . ), d'une part, et pour permettre a tous les visiteurs de réagir sur le sujet évoqué, en postant leurs commentaires sur les articles, créant ainsi une relation privilégiée entre l'auteur et ses lecteurs. Les plateformes de blogs les plus connues sont Overblog 4, Blogger 5, SkyrockBlog 6 et CanalBlog 7.

- **Forum** : Un forum est un lieu d'échange d'informations ou un espace virtuel de négociation et d'échange sur un thème précis et entre différents acteurs qui permet de discuter librement sur plusieurs sujets divers. Les différentes contributions forment un fil de discussion (thread en anglais). Chaque forum de discussion se consacre a un thème précis.

  Les messages publies dans les forums sont archives. Ceci permet aux internautes d'y participer d'une manière asynchrone. Contrairement aux blogs, les messages sont organisés chronologiquement du plus ancien au plus récent.

- **Social bookmarking** : en français marque-page social ou navigation sociale, est un moyen pour stocker, classer, chercher et partager les liens favoris. Ces liens sont accessibles aux utilisateurs d'un site web ou à partir d'un réseau. D'autres utilisateurs ayant les mêmes centres d'intérêt peuvent consulter les liens par sujet, catégorie, étiquette ou même de façon aléatoire. En dehors des favoris Web, on peut trouver d'autres services spécialisés sur un sujet particulier, on cite le site Delicious le plus populaire de social bookmarking.

- **Plate-forme de microblogging**: le microblogging dérivé directement du concept des blogs typiques du web 2.0. La différence réside principalement dans la longueur des publications. Les microbloggeurs sont souvent limites à un nombre de caractères qui est de l'ordre de 140 caractères (cas de Twitter). Toutefois, les microbloggeurs peuvent partager des images ou des liens externes dans leurs messages. Ce facteur encourage par conséquent les internautes a partager des microblogs plus fréquemment.

# I.2 Exemples de réseaux sociaux

## I.2.1 Facebook

Créé par Mark Elliot Zuckerberg, Facebook est à l’origine destiné aux étudiants de l’université d’Harvard (USA). Devant son succès, le site s’est ouvert en septembre 2006 à tous les internautes.

Le nom « Facebook » qui signifie « trombinoscope », vient des photos de classe distribuées en fin d’année scolaire aux étudiants. Le site était donc conçu au départ comme un immense trombinoscope virtuel présentant chaque élève.

Facebook est un réseau social très populaire qui vous permet d'être en lien avec des amis, partager des centres d’intérêts et rejoindre des groupes. Facebook est le plus grand site de réseautage social au monde avec plus de 1.5 milliard d'utilisateurs. L'utilisateur interagit avec en moyenne environ 130 amis sur sa page.


**Figure 2.1 : Page d’accueil Facebook**

Facebook a changé la façon dont nous interagissons sur le web. Les possibilités offertes par le réseau social sont très larges :

- ✓ Lier des liens avec des amis à travers le globe
- ✓ Retrouver des anciens amis de classes
- ✓ Partager des photos, des vidéos avec des amis et sa famille
- ✓ Utiliser des applications pratiques et/ou divertissantes
- ✓ Inviter des amis à un événement
- ✓ Dialoguer avec des amis par messageries instantanées ...

## I.2.2 Google Plus

Google+ est un service à mi-chemin entre Facebook et Twitter. Il nous permet en effet de communiquer avec nos amis en limitant la visibilité de nos messages et photos à un groupe défini de personnes (grâce aux « cercles »). Pour autant, des utilisateurs pourront vous suivre sans que vous ayez besoin de les accepter en tant qu'amis au préalable. Les pages entreprises permettent aux marques de communiquer vers leurs clients.

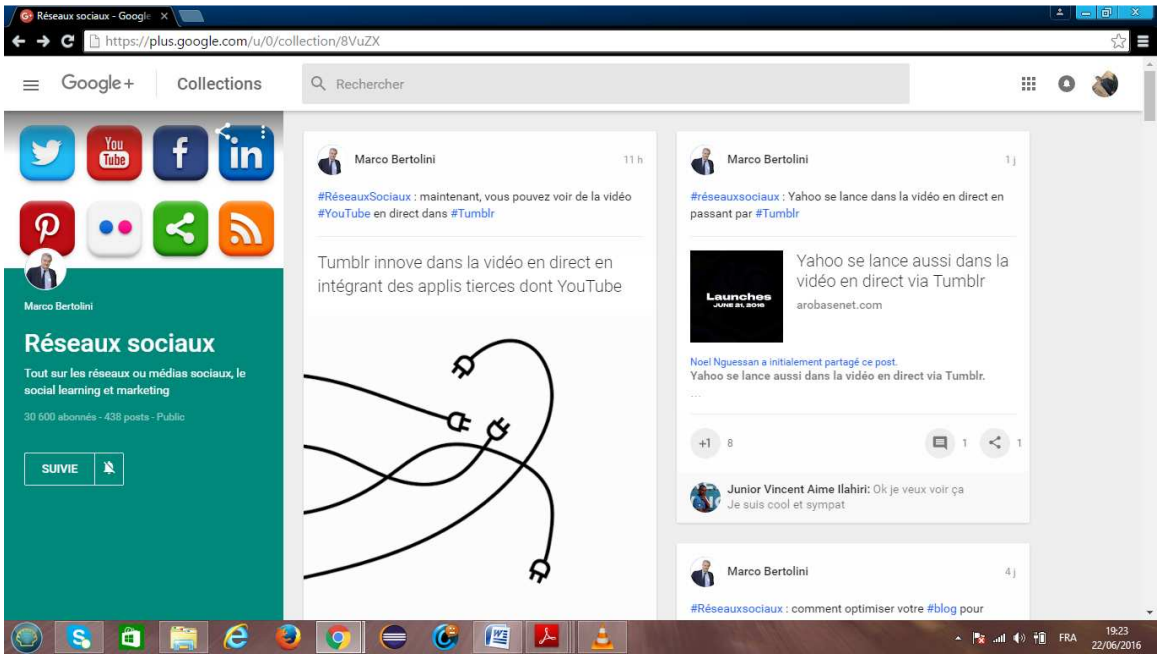


**Figure 2.2 : Page d'accueil Google Plus**

Google+ est l'application de réseau social de l'entreprise américaine Google lancée le 28 juin 2011, et accessible pendant près de 90 jours sur invitation, avant d'être rendue accessible au grand public le 20 septembre 2011. Il est présenté par nombre de médias comme un produit destiné à concurrencer Facebook. Google+ est le deuxième plus grand réseau social au monde, ayant dépassé Twitter en janvier 2013.

Les utilisateurs de Google+ peuvent voir les mises à jour de leurs contacts grâce à des cercles à travers le « Stream », qui est semblable aux « flux de nouvelles » de Facebook. La zone de saisie permet aux utilisateurs de se mettre à niveau sur les états ou l'utilisation des icônes à télécharger et partager des photos et vidéos.

## I.2.3 Linkedin

Linkedin est un réseau social professionnel destiné à faciliter le réseautage entre collègues, clients, partenaires, fournisseurs. Ce réseau social permet également de mettre en valeur son curriculum, de soigner sa réputation, de participer à des groupes de discussion et de faire sa veille d’information en suivant l’actualité de votre secteur d’activité. Enfin, Linkedin est le terrain de chasse favori des commerciaux et cabinets de recrutement en recherche de clients à contacter et de profils qualifiés à placer dans des entreprises.

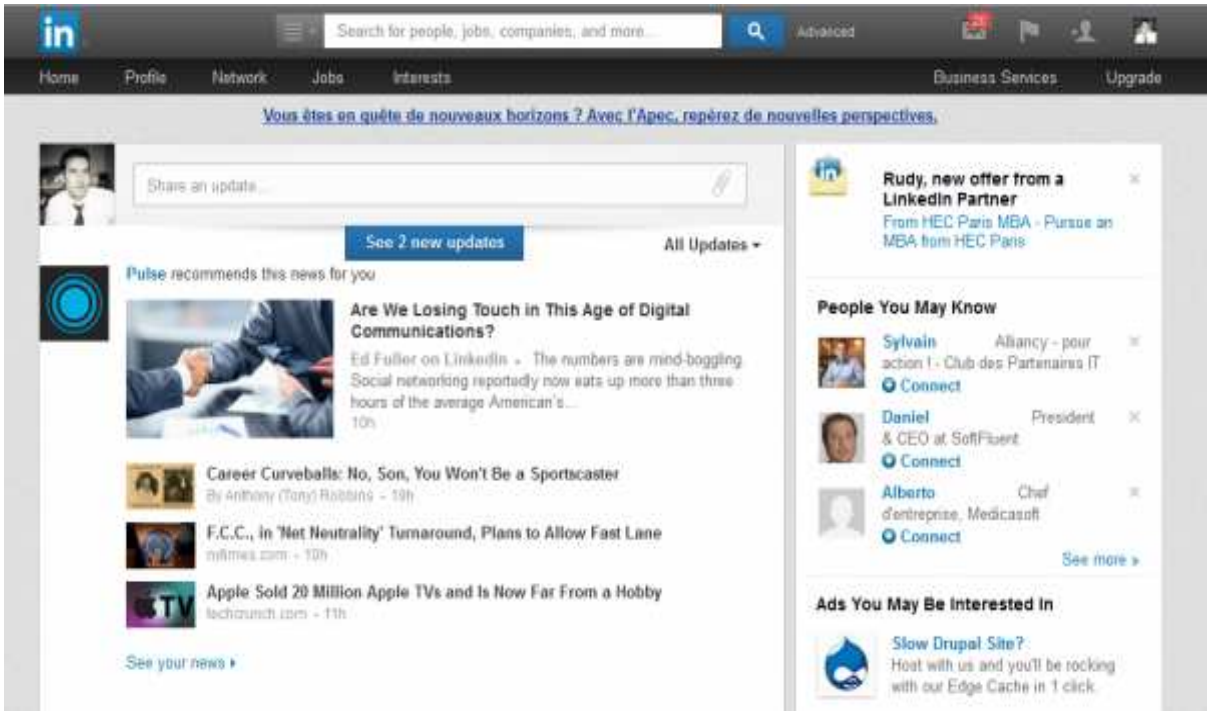


**Figure 2.3 : Page d'accueil Linkedin**

On retrouve aussi le réseau social Twitter que nous avons choisi pour implémenter notre modèle de recherche. Twitter sera décrit ultérieurement en détail.

# II. Les plateformes de microblogging

## II.1 Définition

Une plateforme de microblogging est un système communication et de collaboration qui permet le partage et la diffusion de messages textuels.

## II.2 Exemple de plateformes de microblogging : TWITTER

### II.2.1 Présentation générale de Twitter

Twitter vient du mot **TWEET** qui signifie en anglais "gazouillis"(Le crie d'un oiseau). Twitter est un réseau social de la catégorie de microblogues qui permet d'envoyer des messages de 140 caractères dans lesquels on peut inclure un lien, une image ou même une vidéo mais tout ca sera compté dans le nombre de caractères autorisés.

Chaque message publié possède sa propre adresse URL et est répertorié dans les moteurs de recherche. L'avantage de Twitter est qu'il permet de partager rapidement de l'information sur le web et à avec son réseau.

### II.2.2 Lancement et évolution

Lancé en octobre 2006, la plate-forme de twitter comptait 94,000 utilisateurs en avril 2007 pour atteindre 200 millions en 2012. Au début de 2014 Le nombre d'abonnés de Twitter a atteint les 645 millions.

L'idée de Twitter a été imaginée par Jack Dorsey (PDG de l'entreprise Twitter Inc et fondateur en collaboration avec Evan Williams, Noah Glass, Biz Stone de twitter) lors d'une session de "brainstorming" au sein de la société Odeo, service gratuit de diffusion. En 2006, lorsque Jack Dorsey a imaginé un système qui permettrait aux utilisateurs de décrire ce qu'ils étaient en train de faire sur le moment et de pouvoir le partager via SMS.

Twitter permet aux amis, aux familles et aux collaborateurs de communiquer et de rester connectes en partageant des tweets.

### II.2.3 Créer un compte Twitter

Pour créer un compte Twitter on procède comme suit :

**a)** Saisir l'URL suivante : www.twitter.com. La page d'accueil de twitter s'affiche.

**b)** On vous demande soit de vous connecter, si vous déjà un compte. Sinon, il faut d'abord créer un compte en suivant les étapes ci-dessous :

**Figure 2.4 : Formulaire d'inscription sur Twitter**

**c)** Insérez votre prénom et votre nom dans le champ « Nom complet ».

**d)** Entrez votre adresse email pour que Twitter vous envoie les paramètres d'activation de votre compte twitter ;

**e)** Choisissez un mot de passe sécurisé et le saisir dans le champ : Créez un mot de passe

**f)** Dans le dernier champ « **Choisissez votre nom d'utilisateur »** saisissez un nom qui vous identifie dans le réseau Twitter ;

**g)** Cliquez sur Créer mon compte

**h)** Accéder à votre boite email (que vous avez saisi en **d)**), vous allez trouver un message envoyé depuis twitter pour activer votre compte ;

## II.2.4 Le vocabulaire spécifique de Twitter

Voici un petit récapitulatif des principaux mots et signes utilisés dans Twitter :

- **Tweet :** un tweet est un message posté sur Twitter. Ce message ne peut pas excéder les 140 caractères, espaces compris. Sur Twitter, lors de la frappe de vos tweets, vous verrez un petit compteur au-dessus de votre message diminuer pour vous avertir du nombre de caractères restants.

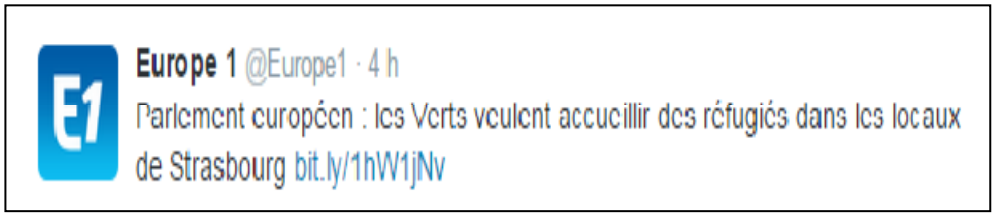


**Figure 2.5 : Exemple d'un Tweet.**

- **ReTweet(RT) :** permet de rediffuser un message d'un autre utilisateur à vos abonnés. Le message est constitué comme tel : **RT @auteurdutweet** message.
- **Follower**(abonné) : c'est une personne qui a décidé de suivre votre file d'actualités.
- **Following(abonnement)** : c'est uen personne que vous, vous avez décidé de suivre.
- **Mention(@)**: comme son nom l'indique, elle permet de mentionner quelqu'un dans un tweet. La mention s'exprime par le symbole **@** accolé à un pseudo, par exemple : « Bonne journée **@ALI** ». Ce message est public, il sera vu par l'ensemble de vos followers.
- **Hashtag(#)** : vient de l'anglais "hash" signifiant "dièse", et "tag" signifiant "mot". C'est donc un moyen d'ajouter de l'information pour préciser, catégoriser un de vos tweet selon un contexte particulier ou pour le lier à d'autres tweets.
- **Direct Messages :** lorsque deux utilisateurs qui se suivent mutuellement souhaitent discuter de manière privée entre eux, ils peuvent utiliser la messagerie privée.
  Il est désormais possible d'activer la fonctionnalité « Recevoir des messages privés de la part de n'importe quel abonné » afin de ne plus être obligé de suivre un utilisateur pour pouvoir communiquer.
- **Follow friday(#FF)**: permet de citer des comptes que vous jugez particulièrement intéressant, l'une des activités du vendredi.
- **Timeline :** elle est l'équivalent du mur Facebook. Il regroupe l'ensemble des tweets et conversations publiées par les utilisateurs que vous suivez. Contrairement à Facebook, la Timeline n'est pas régie par un algorithme (EDIT : bien que désormais Twitter vous suggère des tweets intéressants de personnes que vous ne suivez pas) et les tweets sont affichés au fur et à mesure de leur publication dans un flot continu d'information, permettant ainsi de suivre l'actualité en temps réel. C'est aussi depuis la Timeline que l'utilisateur peut poster un tweet qui sera reçu par l'ensemble de ses abonnés.

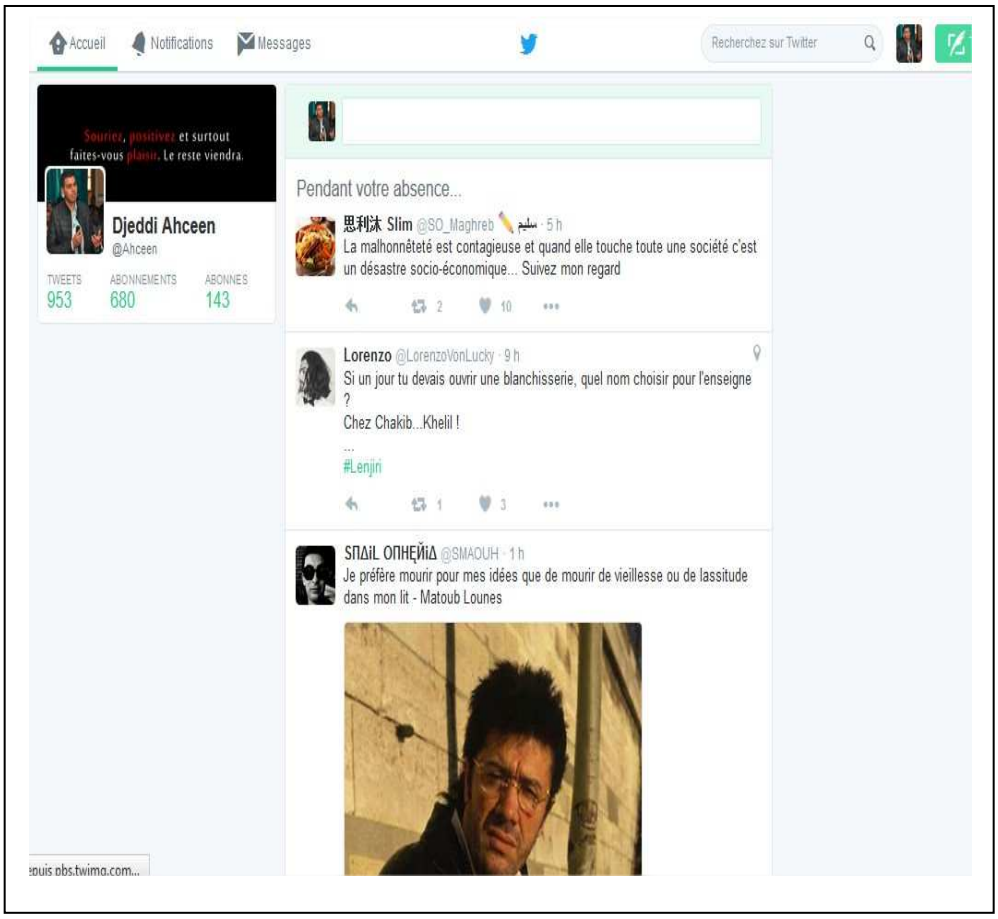


**Figure 2.6 : Exemple de Timeline de Twitter**

- **Tendances** : les tendances sont le fruit de l'activité en temps réelle sur Twitter. Plus les utilisateurs parlent d'un même sujet, plus il aura de chance de se retrouver dans les tendances du jour.

## II.2.5 Le contenu d'un tweet

- Un tweet contient en moyenne 15 mots pour 140 caractères [Jansen et al.,2009b]. La taille d'un tweet est extrêmement faible comparé aux autres sources

d'informations.

- En plus du texte, un tweet peut contenir différent types de signe(#, @, RT, DM,etc).
- Un tweet peut également contenir des liens hypertextes. Ces liens prennent une forme réduite en raison du nombre de caractères autorisés.
- Les Bloggeurs peuvent mettre différents types de multimédias (images, vidéos, audio).
- Dans le réseau de Twitter, les relations d'abonnement peuvent être dans un seul sens, mais également dans les deux sens si B s'abonne à son tour à A.
- Le partage d'information se fait en temps réel.
- Les microblogs contiennent également des métadonnées de différentes natures :
    - de géolocalisation : ces informations permettent de localiser, grâce au GPS, l'endroit duquel le microblog a été publié.
    - d'horodatage: chaque microblog est caractérisé par sa date de publication.
    - d'auteur: les plate-formes de microblogging stockent le compte depuis lequel est publié chaque microblog. Ceci permet aux utilisateurs de trouver les microblogs d'un auteur en particulier.
    - de rediffusion : dans Twitter, on peut connaître le nombre de fois qu'un tweet a été retweeté. On peut également accéder à la liste des utilisateurs qui ont retweeté un tweet donné.

### II.2.6 Réseau social d'information de Twitter : principales entités et relations

- **Les blogueurs** sont les principaux acteurs dans un service de microblogage. Ils sont à la fois consommateurs et producteurs d'information. Les blogueurs sont reliés les uns aux autres au moyen des relations d'abonnement. Une telle relation permet à un utilisateur d'indiquer sont intérêt à d'autres blogueurs et de suivre leurs flux de tweets.

Au cours de son activité de microblogging, un blogueur interagit avec d'autres entités telles que les tweets, les retweets, les réponses, les hashtags et les ressources Web. L'ensemble de ces entités forme *le réseau social d'information*.

- **Les tweets** représentent les principales entités d'information dans le réseau. Les tweets sont visibles par défaut à tous les utilisateurs. Cependant, le blogueur peut restreindre l'accès à ses tweets seulement à ses abonnés.

- **Les retweets** sont des tweets transmis à leur tour par un blogueur à ses propres abonnées. Un retweet maintient toujours une référence vers son auteur d'origine. On distingue deux types de retweets sur Twitter.

- **Les réponses** sont des tweets envoyés à un utilisateur particulier. L'identifiant de destinataire est dans ce cas mentionné au début de la réponse « **@nom** ». Les mentions permettent également d'indiquer les blogueurs concernés par un tweet.
- **Les hashtags** sont des termes marquées avec « **#tag** ». Un hashtag permet d'annoter le contenu d'un tweet.

La figure suivante résume les principales entités impliquées dans le réseau social d'information de Twitter et les diverses relations qui les associent.

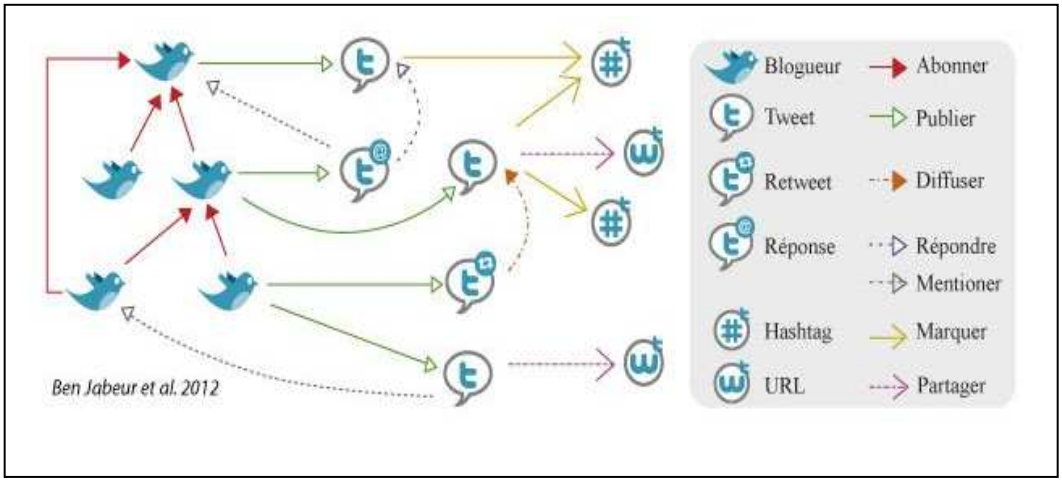

**Figure 2.7 : Le réseau d'information de Twitter**

# III. La recherche d'information sociale (RIS)

L'explosion des réseaux sociaux a conduit à la naissance d'une nouvelle branche de la Recherche d'Information (RI) : la RI sociale. Il s'agit d'adapter les modèles et les algorithmes de la RI classique afin d'exploiter les informations sociales propres à ce nouveau cadre.

***RIS = Recherche d'information classique + Réseaux sociaux***

## III.1 Processus de recherche d'information sociale

Le processus de recherche d'information sociale passe par les mêmes étapes que dans le cas d'une recherche classique avec intégration des informations sociales.

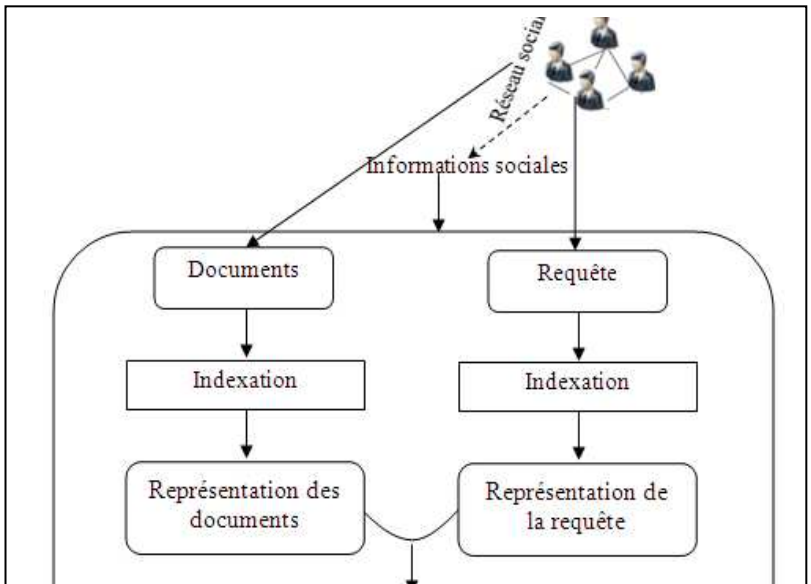

**Figure 2.8 : Processus de RIS**

## III.2 Les informations sociales

Les informations sociales présentent une nouvelle source d'informations. Elles permettent de décrire les utilisateurs et les ressources. On distingue deux types d'information (Damak, 2014) :

- **Contenus générées par les utilisateurs (UGC)** : avec l'arrivée des RS, les utilisateurs contribuent à leur tour à la production d'informations. Leurs contributions peuvent être de différentes natures tels que :
  - les contenus publiés dans les plate-formes sociales telles que les blogs et les wikis,
  - les réactions,
  - les informations publiées par les autres utilisateurs telles que les annotations et les commentaires,
  - etc.
- **Contenus générées par la pratique :** c'est l'ensemble des informations produites par les utilisateurs lors de :
  - les traces des utilisateurs : elles comportent les différentes pages web visitées par les utilisateurs, les clics, les durées de visites. . . Ces données
  - peuvent être utilisées afin de déterminer les préférences des utilisateurs et leurs thématiques de recherche.

- les données personnelles : elles se composent des informations que l'utilisateur fournit au moment de son inscription sur les réseaux sociaux.
- les liens sociaux : ces liens peuvent être :
    - Symétriques : dans les réseaux tels que Facebook on établi une relation directe entre les personnes que nous avons contacté et qui s'appellent "amis", c'est à dire : pour que je soit ami d'une personne quelconque cette personne doit être mon ami, il s'agit donc d'une relation symétrique ou réciproque.
    - Asymétriques : dans Twitter par exemple, suivre une personne ne l'oblige pas a nous suivre, il s'agit d'une relation asymétrique.

## III.3 Recherche d'informations dans les microblogs: cas de Twitter

La recherche d'information dans les microblogs est particulièrement limitée par la taille courte des articles qui augmente à son tour la difficulté de la recherche textuelle par mots-clés.

Les approches de RI classiques, élaborées pour traiter les documents traditionnels ou des documents de type page Web et qui se basent principalement sur le contenu textuel des documents et sur des statistiques des fréquences de termes, ne sont plus adaptées aux spécificités des microblogs. Les travaux dans ce domaine tentent d'adapter les modèles et techniques de recherche de la RI classique pour la RI dans les microblogs. Ces approches peuvent être classées comme suit :

### III.3.1 Recherche d'information temps-réel

L'un des avantages des microblogs en général et de Twitter en particulier et son caractère temps-réel et l'indexation automatique des Twittes publier, ces avantages qui

concordes avec les besoins des utilisateurs de l'information la plus fraîche possible et combler le vide que les autres sources de web laisse à cause du temps plus long qu'ils leurs faut pour indexer les documents.

(Ounis et al.2011) propose une approche qui commence par faire un classement selon les facteurs temps avant de faire la recherche selon les autres facteurs ce qui revient à classer anti-chronologiquement les résultats puis à éliminer ceux qui sont de faible pertinence.

## III.3.2 Recherche d'opinions et des tendances

- **Opinion**

Les microblogs sont les principales plateformes où les gens expriment leurs opinions sur divers sujets. Des chercheurs se sont intéressés à la problématique comment analyser et exploiter ce type d'informations.

(Bollen et al, 2009) ont utilisé un enregistrement d'événements populaires recueillies auprès des medias puis, ils ont comparé chaque jour avec un vecteur d'humeur à six dimensions extrais à partir du contenu publié sur Twitter , les états d'humeurs utilisent un instrument psychométrique et les états sont (la tension, la dépression, la colère, la vigueur, la fatigue, confusion) puis modélise les tendances collectives émotives en fonction des prédictions faites sur l'analyse des opinions dans les réseaux sociaux et des indicateurs économiques (bourse, prix de matière premiers… etc.)

- **Tendance**

La détection de tendances vise à identifier automatiquement les thèmes émergeant qui apparaissent dans le flux de microblogs en temps-réel.

En marketing par exemple, l a détection de tendance est très importante. Elle a pour but de détecter et anticiper les opinions des gens et surtout d'identifier avec certitude et précision cette tendance le plus rapidement possible, afin de pouvoir réagir(adapter l'offre de l'entreprise à la tendance).

### III.3.3 Recherche des microblogueurs

La recherche de microbloggeurs s’apparente à la tâche de recherche d’experts de la RI classique. Les objectifs sont l’identification des utilisateurs les plus populaires, ceux qui ont les mêmes centres d’intérêts que l’utilisateur courant, ou bien les experts dans des domaines spécifiques. Une approche proposée par (Li et al, 2012b) vise à déterminer les utilisateurs de Twitter qui publient en premier les informations sur un sujet donné.

# Conclusion

Nous avons présenté tout au long de ce chapitre tout d’abord les microblogs , leurs particularités et la particularité des documents qu’il génèrent. Puis nous avons présenté la recherche d’informations sociale en parlant tout d’abord des différentes informations sociales générées par les microblogs, puis nous avons classifié les approches proposés pour adapter les modèles de la RI classique à la RI sociale.

# Chapitre III :

# Intégration du facteur temps dans la RI dans les microblogs

# Introduction

Comme nous l'avons vu dans le chapitre précédent, la recherche d'information sociale consiste à exploiter les signaux sociaux et les intégrer comme facteurs de pertinence dans les modèles de recherches.

Nous avons aussi vu que les utilisateurs qui effectuent des recherches sur twitter sont motivés en premier lieu par la recherche d'information récente, ce qui nous fait dire, que le temps est un facteur important pour les utilisateurs.

Dans ce présent chapitre, nous présentons d'abords les travaux qui se sont intéressés au facteur temporel dans les microblogs pour ensuite présenter nos deux approches qui intègrent le temps de deux manières différentes.

# I. Etat de l'art

L'utilisateur cherche à avoir l'information la plus récente, et pertinente, par rapport à un besoin d'information. Plusieurs travaux ont exploité le facteur temps pour mettre au point des modèles de recherche dans les microblogs dont nous citons:

- L'approche de (Massoudi et al. 2011) qui vise à calculer un score de fraicheur basé sur la déférence de temps entre la date de création du document et la date de la soumission de la requête. Cette approche permet de classer les résultats du plus ancien vers le plus récent, de manière a favorisé les tweets les plus récents.

- L'approche proposée par (Efron et al, 2012) vise à construire un profil pour chaque microblog. Ce profil est caractérisé par les microblogs qui discutent du même sujet ainsi que les périodes de temps pendant lesquelles ont été publiés ces microblogs. Puis le profil est utilisé pour générer et choisir l'information supplémentaire, qui sera utilisé pour enrichir la représentation du microblog ou bien de la requête.

- (Lin et al, 2012) ont proposé (TASE : Time-AwareSearch Engine) un moteur

de recherche qui exploite le facteur temps en calculant les relations entre la date de publications des documents et leurs similarités thématique. Cette approche est inspiré des travaux de la littérature (Kanhabua et Nørvåg, 2011) donc elle combine plusieurs facteurs de pertinences (thématique, temps, pertinence sociale) de façon linaire, avec des coefficients d'amortissements pour chaque facteur.

- (Damak, 2014) propose d'intégrer le temps de différentes manières :
    - Dans un premier temps, il propose d'amplifier les scores de pertinence du contenu d'un tweet en fonction de sa proximité temporelle avec la date de la requête.
    - Dans un deuxième temps, Damak propose de favoriser les termes fréquemment utilisés au moment de la soumission de la requête.

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

- Dans cette troisième méthode, Damak prpose d'amplifier le score d'un terme dans un tweet publié à un instant *t* en fonction de la fréquence d'emploi de ce terme dans cette période *t*. Un même terme aura des scores différents en fonction de la date de soumission du document auquel il appartient. Ce score sera plus important si le terme appartient à un document publié dans une période de rafale de ce terme, que dans le cas où il appartient à un document publié dans une période où le terme n'est pas fréquemment utilisé.

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

# II. Approches proposées

Nous proposons dans ce qui suit deux approches pour la recherche sociale des *tweets* qui associent la pertinence des articles (pertinence thématique) à l'importance sociale des Tweets correspondants (pertinence sociale). Ces deux modèles tentent d'exploiter le facteur temps et le nombre de retweets pour déterminer l'importance sociale d'un tweet.

# II.1 Approche 1

**a) Principe :** un blogueur confirme par la retransmission d'un Tweet, l'importance du message communiqué. L'importance d'un *Tweet* est alors déterminée par le nombre de fois qu'il a été retransmis (nombre de retweets) dans une période de temps.

La figure ci-dessous nous montre deux tweets qui sont retransmis tous les deux 21 fois sauf que le Tweet 1 a pris 11 h pour atteindre ce nombre contrairement au Tweet 2 qui n'a pris que 9h. De là, nous pouvons conclure que le Tweet 2 est plus influent que le Tweet 1. De là nous est venu l'idée de déterminer l'importance d'un Tweet selon sa **vitesse de retweet**.

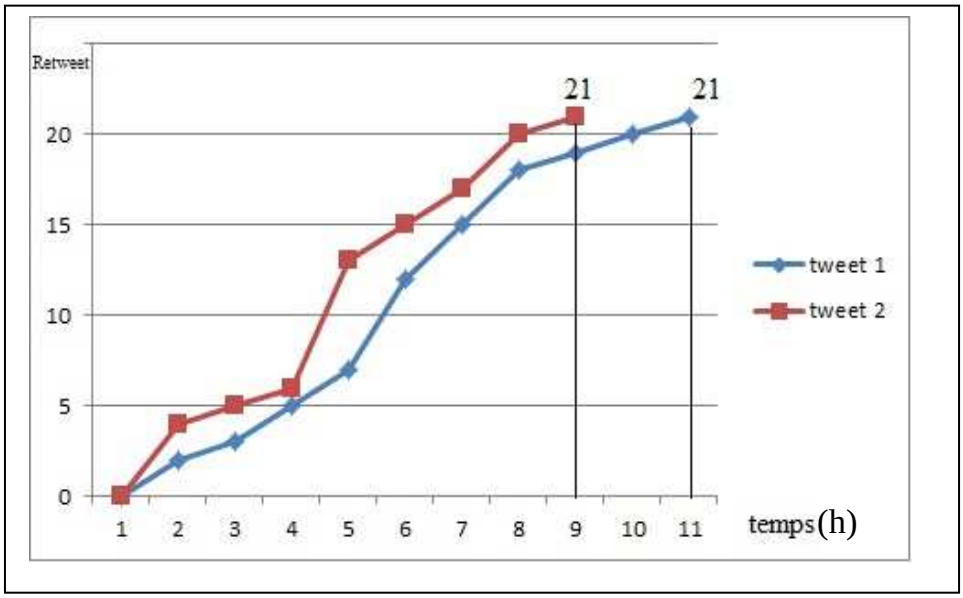


**Figure3.1 : Exemple de deux Tweets avec le même nombre de RT**

**b) Formulation du modèle proposé :** notre modèle combine un score de pertinence thématique et un score de pertinence social du Tweet. L'objectif de cette combinaison est de présenter une liste de tweets qui couvrent le sujet de la requête et qui ont une vitesse de retransmission plus grande. Ces deux scores sont combinés linéairement selon la formule suivante :

$$Score\ (Q, Ti) = \alpha * Score_{Thématique}\ (Q, T_i) + \beta * Score_{Social}\ (T_i, D, Nbr\_RT)$$

Où ***Q*** et ***Ti*** représentent respectivement la requête et le Tweet.

***D*** et ***Nbr_RT*** représentent la durée et le nombre de retweets respectivement.

***α et β*** sont des coefficients d'amortissements, obtenus par expérimentations de manière à avoir les résultats les plus optimaux

La pertinence thématique dépend uniquement du Tweet et de la requête. Pour calculer le score thématique nous avons utilisé le modèle vectoriel de LUCENE. Concernant l'importance sociale $Score_{Social}$, nous précisons que ce score prend en compte le nombre de retweet du tweet ainsi que la durée de retweet (c'est le temps ***D*** pris par le tweet pour atteindre un nombre de retweets ***Nbr_RT***). Dans la suite, nous détaillons le calcul de ce score.

**c) Calcul du score social :** le score sociale $Score_{Social}\ (T_i, D, Nbre\_RT)$ évalue l'importance d'un Tweet $T_i$ en fonction de son nombre de retweets ***Nbre_RT et*** la durée de retweets ***D.***

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

Où $\mu_{Nbr\_RT}$ est une valeur qui permet de privilégier les tweets qui possèdent un taux élevé de retweets. Par expérimentation, nous avons décidé que $\mu_{Nbr\_RT} = 0,5$ si le nombre de retweets**, Nbr_RT,** est supérieur à 50 sinon est $\mu_{Nbr\_RT} = 0$.

$Score_{Vitesse\_RT}\ (T_i, D, Nbr\_RT)$ permet de calculer la vitesse de retweet du Tweet $T_i$. Ce Score est le rapport entre le nombre de retweets ***Nbr_RT*** et le temps qu'il fallut au Tweet pour atteindre ce dernier ***D***. Il est donné par la formule suivante :

$$Score_{Vitesse\_RT}\ (T_i, D, Nbr\_RT) = Nbr\_RT / D$$

La durée ***D*** est déterminée de la manière suivante :

$$D = \text{Date du dernier retweet du Tweet } T_i - \text{Date de création du tweet } T_i$$

# II.2 Approche 2

**a) Principe :** après avoir testé l'approche 1, définie précédemment, nous avons constaté que la vitesse de retweet ainsi définie n'a pas donné de résultats satisfaisants (à voir dans le chapitre suivant). Nous avons eu l'idée de calculer cette même vitesse d'une autre manière.

Pour calculer la vitesse des retweets, nous procédons comme suit :

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

Ce procédé nous permet d'obtenir une vitesse de retweet qui prend en compte toute la durée des retweets, de la date du 1er retweet à la date du dernier retweet, ce qui garantit un traitement égale pour les tweets influant sur une longue durée. Ce qui permet de ne pas être fossé pas les passages à vide.

L'idée de cette approche nous est venue de la manière dont on calcule la précision moyenne dans la recherche d'information classique.

**b) Formulation du modèle proposé :** les deux scores de pertinence thématique et sociale sont combinés linéairement de la même manière que celle définie en approche 1.

$$Score(Q, Ti) = \alpha * Score_{Thématique}(Q, T_i) + \beta * Score_{Social}(T_i, D, Nbr\_RT)$$

Où ***Q*** et ***Ti*** représentent respectivement la requête et le Tweet.

***D*** et ***Nbr_RT*** représentent la durée et le nombre de retweets respectivement.

***α et β*** sont des coefficients d'amortissements, obtenus par expérimentations de manière à avoir les résultats les plus optimaux.

Pour calculer le score thématique nous avons utilisé le modèle vectoriel de LUCENE (de la même façon que l'approche 1). Concernant l'importance sociale $Score_{Social}$,

c) **Calcul du score social :** le score sociale $Score_{Social}(T_i, D, Nbre\_RT)$ évalue l'importance d'un Tweet $T_i$ en fonction de son nombre de retweets ***Nbre_RT*** la durée ***D*** de retweet numéro ***j*** ainsi que son rang $R_i$.

$$Score_{Social}(T_i, D, Nbr\_RT) = Score_{Vitesse\_RT}(T_i, R_{ij}\ D_{ij}, Nbr\_RT_{ij})$$

Avec ***Ti*** le tweet, ***Rij*** le retweet de $T_i$ numéro ***j***, $D_{ij}$ la durée de retweet numéro ***j*** du tweet $T_i$ et $Nbr\_RT_{ij}$ le classement du retweet donc le nombre de retweets à l' instant ***j***.

Le score $Score_{Vitesse\_RT}$ est calculé comme suit :

$$Score_{Vitesse\_RT}(T_i, R_{ij}\ D_{ij}, Nbr\_RT_{ij}) = (\sum (Nbr\_RT_{ij} / D_{ij})) / Nbr\_RT$$

La durée est calculée comme suit :

$$D_{ij} = Date\ du\ retweet\ numéro\ j\ du\ Tweet\ T_i - Date\ de\ création\ du\ tweet\ T_i$$

# Conclusion

Nous avons présenté dans ce chapitre, nous avons décrit en détail les deux approches que nous proposons. Ces deux approches tentent d'exploiter le facteur temps et le nombre de retweets, de deux manières différentes, afin d'améliorer le score thématique d'une recherche dans les microblogs de Twitter.

Dans le prochain chapitre, nous allons expérimenter ces deux approches afin de voir s'ils rapportent un plus à la recherche thématique.

# Chapitre IV :
# Evaluation expérimentale

# Introduction

Dans ce présent chapitre, nous décrivons les différents environnements et outils utilisés afin d'implémenter les deux modèles présentés dans le chapitre 3. Ensuite, nous décrivons le cadre expérimentale de nos approches, pour terminer par l'affichage des résultats obtenus et les discuter.

# I. Outils de développement

## I.1 Eclipse IDE

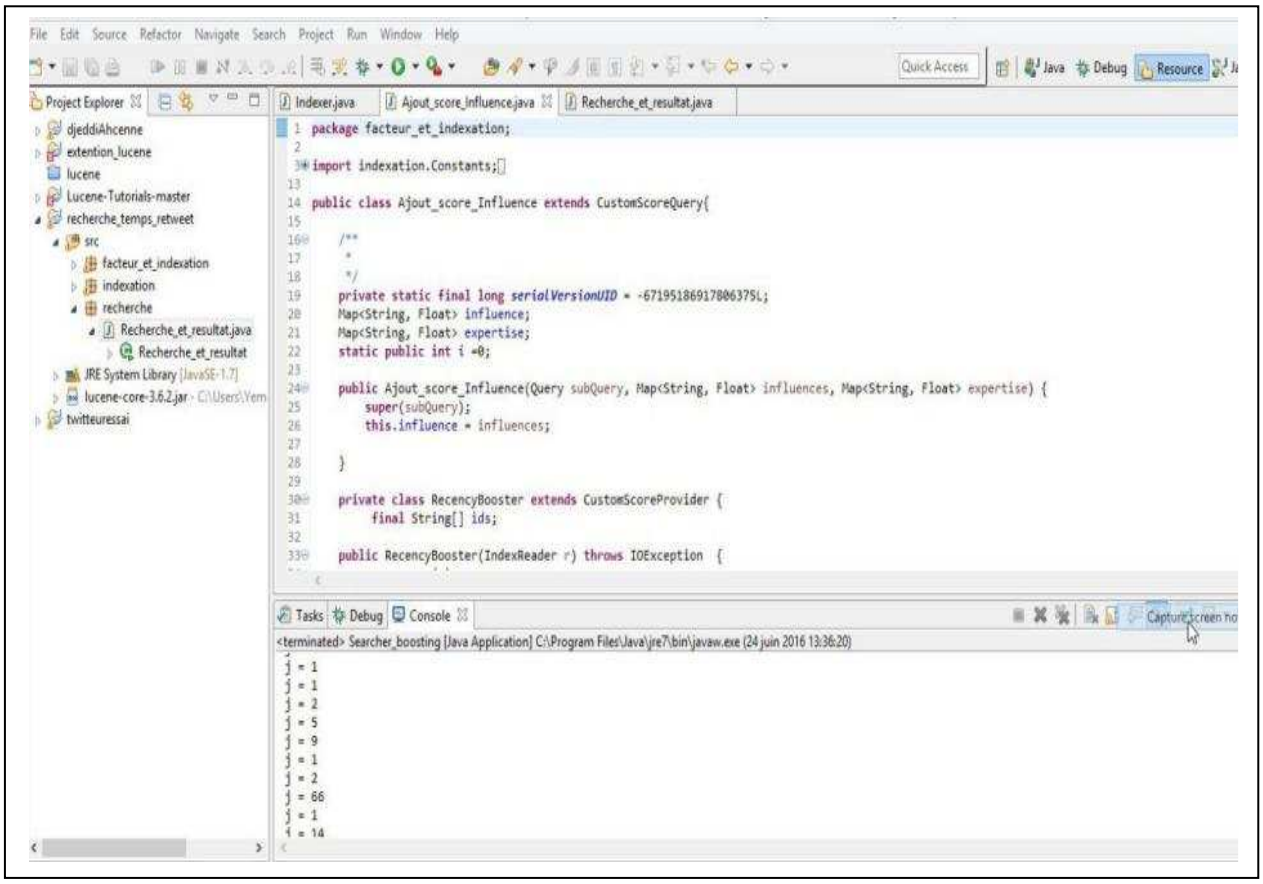


**Figure 4.1 : Interface d'Eclipse**

Dans un environnement de développement « intégré » (abrégé EDI en français ou IDE en anglais, pour Integrated Development Environment) les outils sont prévus pour être utilisés ensemble (le produit d'un outil peut servir de matière première pour

un autre outil) . Les outils peuvent être intégrés dès le départ, c'est-à-dire qu'ils sont construits dans le but d'être utilisés ensemble. Il peut aussi s'agir d'un ensemble d'outils développés sans lien entre eux et intégrés à posteriori.

**Eclipse** est un environnement de développement IDE. Il est gratuit et disponible pour la pluparts des systèmes d'exploitation qu'on peut trouver sur le marché.

## I.2 Le langage JAVA

Le langage **Java** a été présenté officiellement le 23 mai 1995 au SunWorld**.** Il s'agit d'un langage de programmation informatique orienté objet créé par James Gosling et Patrick Naughton, employés de Sun Microsystems.

Le langage Java reprend en grande partie la syntaxe du langage C++, mais Java ne posséde pas certains concepts de ce dernier, tels que les pointeurs et les références, ou l'héritage multiple contourné par l'implémentation des interfaces. Les concepteurs ont privilégié l'approche orientée objet de sorte qu'en Java, tout est objet à l'exception des types primitifs (nombres entiers, nombres à virgule flottante, etc.).

Les programmes écrits dans en Java ont la faculté d'être facilement portables sur différents systèmes d'exploitation (UNIX, Linux, Windows ...etc.), avec peu ou pas de modifications. Pour cela, divers plateformes et Framework sont associés afin de garantir cette portabilité qui est l'un des objectifs principaux de ce langage**.**

Java possède plusieurs versions. Pour implémenter notre approche nous avons choisi la version ***Java SE 8*** qui est sortie en Mars 2014.

## I.3 LUCENE

Lucene est un projet de la fondation Apache. Au début il était mis sous licence GPL et maintenant sous licence Apache. Il s'agit d'une bibliothèque open source écrite en java qui comprend les tâche principales d'un moteur de recherche ce qui veux dire qu'elle possède les classes nécessaires pour l'indexation et l'appariement requêtes/documents. Parmi les nombreuses classes contenues dans LUCENE nous citons :

- **Les Classes d'indexation**

  o **IndexWriter** : cette classe est le composant central du processus d'indexation. Elle permet de créer un nouvel index et ajouter des documents à un index existant. On peut se la représenter comme un objet par lequel on peut écrire dans l'index mais qui ne permet pas de le lire ou de le rechercher.

  o **Analyzer** : avant que le texte soit dans l'index, il passe par l'Analyser. Celui-ci est une classe abstraite qui est utilisée pour extraire les mots importants pour l'index et supprime le reste.

  Cette classe tient une part importante dans Lucene et peut être utilisée pour faire bien plus qu'un simple filtre d'entrée.

  o **Document** : la classe Document représente un rassemblement de champs. Les champs d'un document représentent le document ou les métadonnées associées avec ce document. La source originelle (comme des enregistrements d'une base de données, un document Word, un chapitre d'un livre, etc.) est hors de propos pour Lucene.

  Les Métadonnées comme l'auteur, le titre, le sujet, la date, etc. sont indexées et stockées séparément comme des champs d'un document.

.

- ***Classes de recherche***

  o **IndexSearcher :** la classe IndexSearcher est à la recherche ce que IndexWriter est à l'indexation. On peut se la représenter comme une classe qui ouvre un index en mode lecture seule.

  o **TermQuery** : c'est la méthode la plus basique d'interrogation de Lucene. Elle est utilisée pour égaliser les documents qui contiennent des champs avec des valeurs spécifiques.

  o **QueryParse**r : la classe QueryParser est utilisée pour générer un décompositeur Analytique qui peut chercher à travers un index.

## I.4 Twitter4J

L'API Twitter4j est une bibliothèque écrit en Java qui permet de récupérer des informations sur twitter. C'est aussi une passerelle ou interface de programmation permettant de se connecter aux données Twitter de façon automatisée.

L'API Twitter4j peut aussi être utilisée pour afficher automatiquement des tweets sur un site web ou pour extraire des données à des fins de veille sur les réseaux sociaux. Dans ce dernier cas on utilise la Search API de Twitter et on peut par exemple l'utiliser pour extraire et faire défiler à l'écran les tweets utilisant le hashtag lié à une émission de télévision.

# II. Démarche d'évaluation

## II.1 Description de la collection de test

Pour réaliser nos tests, nous avons utilisé une collection gratuite, disponible sur le net. Cette collection contient

- 502 tweets qui se situent entre Mai 2010 et Mars 2011.
- 802 requêtes.
- 802 jugements de pertinence.

## II.2 Mesures d'évaluation

Pour évaluer les deux modèles définis précédemment, nous avons opté à utiliser les mesures d'évaluations standards, à savoir :

- La MAP
- La R-Précision
- La précision@X

Nous avons aussi utilisés les mesures de base Rappel, précision et la F-mesure.

## II.3 Expérimentations et résultats

Dans cette partie, nous présentons en premier les résultats obtenus lors des expérimentations de la première approche (Approche 1). Ensuite, nous présentons les résultats des expérimentations de l'approche 2.

- **Approche 1**

  - ***Précision@X***

| Score | P@5 | P@10 | P@15 |
|---|---|---|---|
| Thématique | 0,1708 | 0,0888 | 0,0601 |
| Thématique + Sociale | 0,1700 | 0,0892 | 0,0599 |

**Tableau 4.1: P@5, P@10 et P@15 de l'approche 1**

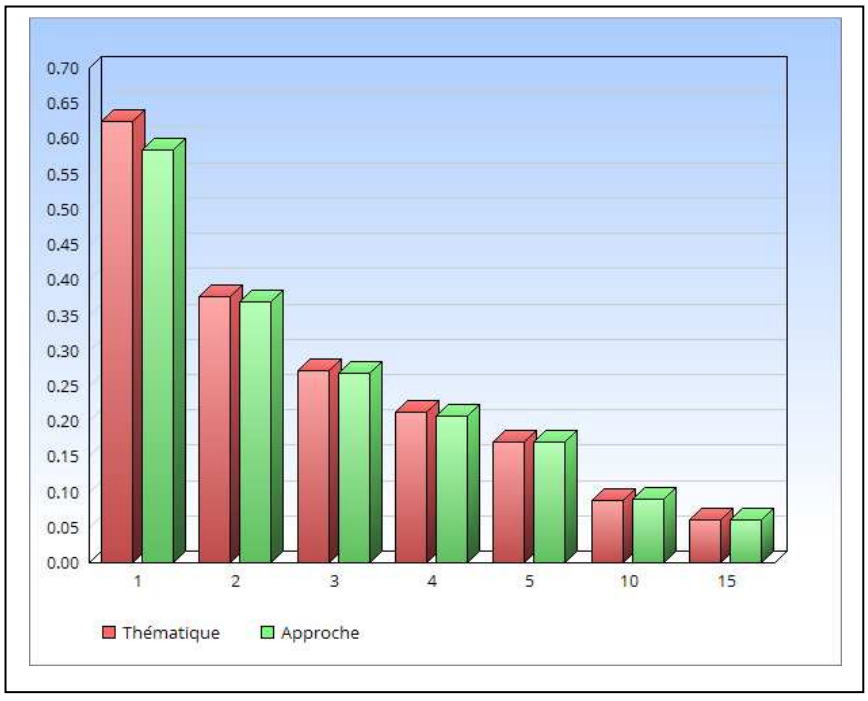


**Figure 4.2 : Comparaison de la Précision@X du score thématique et du score de l'approche 1**

Ces expérimentations nous montrent que, plus le nombre de documents retournés par le système augmente, plus la proportion des documents pertinents

trouvés diminue. Cela montre que le modèle de recherche perd en précision à mesure que le nombre de documents augmente.

- ***Les mesures : R-Précison, MAP et la MAP pour 30 requêtes***

| Mesure d'évaluation | Thématique | Approche I |
|---|---|---|
| **R-précision** | **0.6240** | **0.5837** |
| **MAP** | **0.7223** | **0.6984** |
| **MAP pour 30 requêtes** | **0,2451** | **0,2116** |

**Tableau 4.2 : R-précision, MAP et MAP30 de l'approche 1**

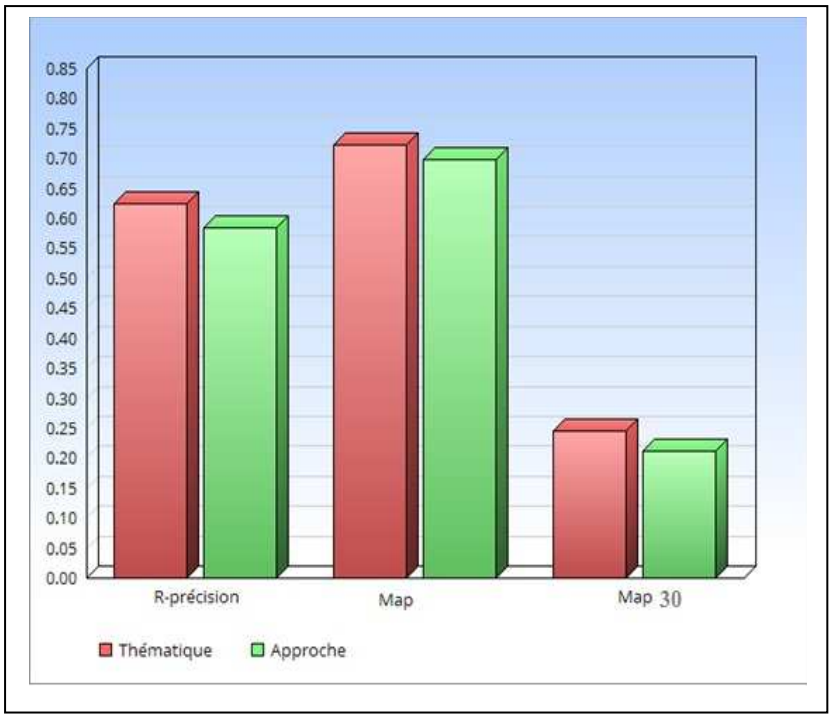


**Figure 4.3: Les résultats des mesures standards de l'approche 1**

Le tableau montre que notre approche n'améliore en aucun cas les résultats par rapport au modèle thématique. Nous avons même constaté une dégradation sensible des résultats. La MAP du score thématique est plus importante que celle de notre approche de 0,0239 donc la précision diminue de 3,3 %. Le même constat est fait avec la R-précision et la MAP avec 30 requêtes, la R-précision diminue de 6,5 %.

- ***Rappel, précision et F-mesure***
  - ✓ **Précision :** P = 470 / 3001 = 0,1566
  - ✓ **Rappel :** R = 470 /526 = 0,8935
  - ✓ **F-mesure :** F-score= (2*R*P)/(R+P) = 0,2996

- **Synthèse :** les nombreuses expérimentations faites sur la collection d'edgard montre que notre approche qui intègre un nouveau score, en plus du score thématique, que nous avons appelé ***vitesse de retweet***, n'importe guère d'amélioration mais au contraire, elle a permis de diminuer la pertinence.

Cette défaillance revient peut être à la petite taille de la collection que nous avons utilisé. Et possible qu'elle soit aussi due aux longues périodes qui séparent les retweets. Cette période peut aller jusqu'à une année.
Le nombre de retweets pour les tweets de la collection utilisée ne varie pas trop, il ne dépasse pas les 100, et la majorité possède un nombre inférieur à 50. Cette particularité peut aussi être la raison de défaillance de l'approche 1.

## ● Approche 2

Pour les différentes expérimentations, nous avons utilisé la même collection utilisée pour l'approche 1, et nous allons utiliser les mêmes mesures d'évaluations.

- ***Précision@X***

| Score | P@1 | P@2 | P@3 | P@4 | P@5 | P@10 | P@15 |
|---|---|---|---|---|---|---|---|
| **Thématique** | **0.6250** | **0.3760** | **0.2724** | **0.2125** | **0.1708** | **0.0888** | **0.0601** |
| **Approche 2** | **0.6327** | **0.3779** | **0.2699** | **0.2115** | **0.1723** | **0.0888** | **0.0599** |

**Tableau 4.3 : Précision @X de l'approche 2**

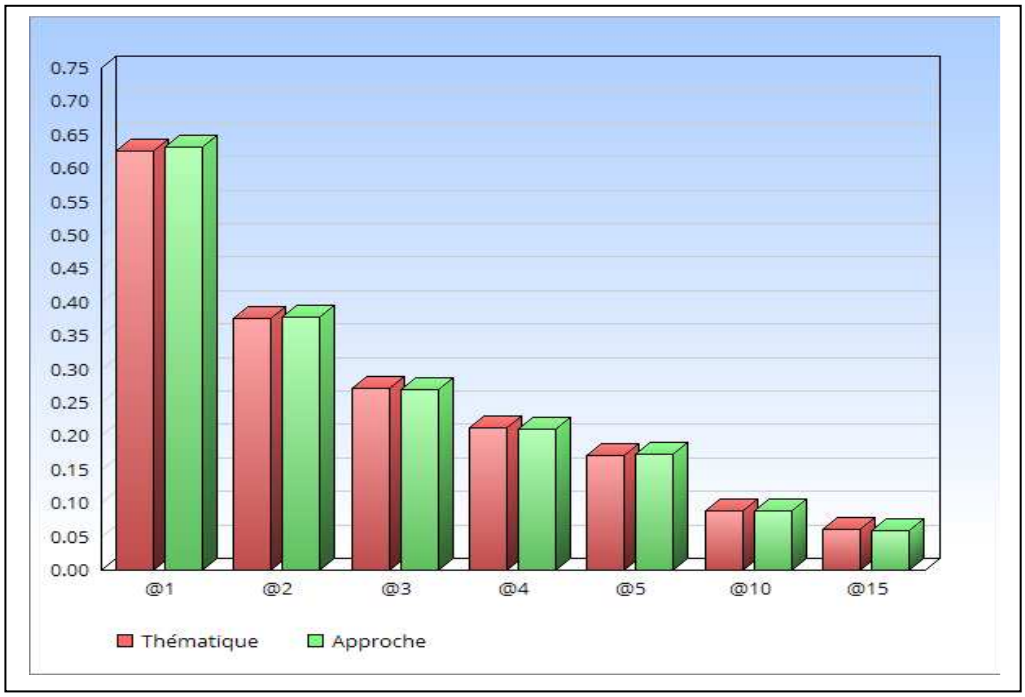


**Figure 4.4: Comparaison de la Précision@X du score thématique et du score de l'approche 2**

Les résultats montrent que notre approche améliore légèrement les résultats. Par exemple, pour les 5 premiers documents retournés il y a une amélioration de 0,0015.

- ***Les mesures : R-Précison, MAP et la MAP pour 30 requêtes***

| Mesure | Thématique | Approche 2 |
|---|---|---|
| **R-précision** | **0.6240** | **0.6317** |
| **MAP** | **0.7223** | **0.7264** |
| **MAP sur 30 requêtes** | **0,2116** | **0,2451** |

**Tableau 4.4: R-précision, MAP et MAP30 de l'approche 2**

Notre approche améliore nettement les résultats selon les métriques utilisés. On voie que notre approche améliore la R-précision de 0,0157(2,4 %). La MAP est améliorée de 0,0041(0,05 %) et la Map30 de 0,0335.

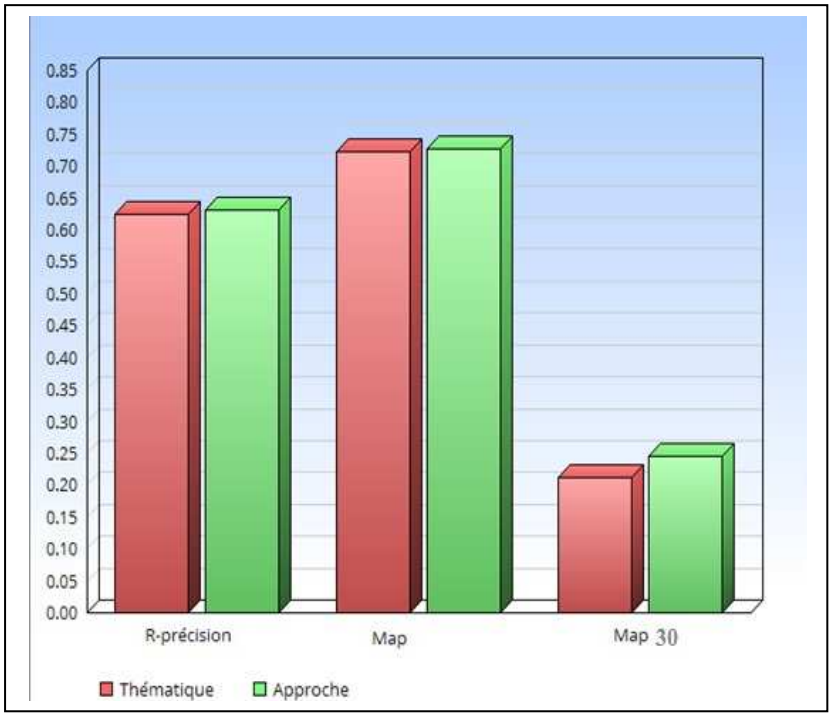


**Figure 4.5: Les résultats des mesures standards de l'approche 2**

- ***Rappel, précision et F-mesure***
  - ✓ **Précision :** P = 470 / 3001 = 0,1566
  - ✓ **Rappel :** R = 470 /526 = 0,8935
  - ✓ **F-mesure :** F-score= (2*R*P)/(R+P) = 0,2996

- **Synthèse :** On peut conclure que cette approche n'améliore pas le rappel ou bien la précision générale, c'est à dire le nombre de documents retournés reste le même. Alors que pour la précision réelle et la MAP qui sont des mesures qui évaluent la vitesse de restitution du système, nous avons constaté une amélioration légère mais qui reste importante.

# Conclusion

Nous avons proposé dans ce dernier chapitre le cadre expérimental des deux approches que nous avons proposé au chapitre 3.

L'évaluation expérimentale que nous avons menée sur une collection de tweets (collection d'Edgard) montre que l'approche 1 n'améliore ne aucun cas les résultats contrairement à l'approche 2 qui nous a donné des résultats assez satisfaisants. De là, nous pouvons conclure que la vitesse des retweets, qui est la combinaison du facteur temps et du nombre de retweets, lorsque elle est exploitée avec le facteur thématique, améliore les performances du système de recherche d'informations à condition de prendre en compte la façon dont les retweets sont distribués dans le temps, de manière optimale.

# Conclusion générale

Notre travail s'est porté sur la recherche d'information sociale dans un contexte de microblogging pour la plateforme et le réseau social twitter. Pour arriver à proposer, formaliser, implémenter et évaluer une approche dans ce contexte nous avons commencé par présenter la recherche d'information classique, nous avons expliqué les déférentes phases de recherche ainsi que les modèles les plus connus qui sont utilisés.

Nous nous sommes intéressés dans ce travail à la recherche d'information adhoc dans les microblogs de Twitter, la plateforme de microblogging la plus populaire. L'objectif est de retrouver les microblogs répondant à un besoin d'information spécifié par un utilisateur. Pour cela, nous avons cherché comment intégrer le facteur temps qui est un facteur crucial dans la recherche d'information.

Nous avons proposé deux approches qui intègrent le temps dans la recherche. Une première approche qui calcule un score de pertinence par la somme d'un score thématique et un score sociale. Ce dernier exploite, parmi les informations sociales, le nombre de retweet et la temporalité du tweet pour évaluer la vitesse de retweet. Mais cette approche n'a malheureusement pas donné de résultats. Alors nous avons pensé à une autre manière de calculer cette même vitesse de façon à prendre en compte les passages à vide, que nous pensons était la cause de l'échec de la première approche.

L'approche 2 est inspirée de la manière dont est calculée la MAP, la vitesse de retweet est calculée à différents passages, chaque fois qu'il y a un nouveau retweet. Cette deuxième approche a apporté des résultats plutôt satisfaisant. Toute fois la taille petite du corpus utilisé n'a pas contribué à avoir une nette amélioration.

# Limites et perspectives

Les deux approches proposées présentent deux limites principales. La première c’est qu'il n’y a pas eu d’améliorations pour ce qui du nombre de documents pertinents retourné, et la deuxième concerne la précision par rapport aux approches basées que sur la thématique, de ce coté aussi, nous n’avons pas eu d’améliorations.

Nous pensons que l'exploitation des autres réactions des utilisateurs sur un tweet (favoris, commentaire …etc.), et l'étude du réseau et des comportements des utilisateurs qui réagissent sur ce tweet, pourraient donner une autre dimension à nos propositions.

# Bibliographie

**[Azzoug .2013]** W. Azzoug:Contribution à la definition d'une approche d'indexation sémantique de documents textuels. Mémoire de Magister, Université Boumerdes, 2013 .

**[Ben Jabeur, L 2013],** Damak, F., Tamine, L., Cabanac, G., Pinel-Sauvagnat, K., et Boughanem, M. (2013). IRIT at TREC Microblog Track 2013. In E. M. Voorhees et (Eds.), *Text REtrieval Conference (TREC), Gaithersburg, USA,*. National Institute of Standards and Technology (NIST).

**[Bollen et al. 2009]** Bollen, J., Pepe, A., et Mao, H. (2009). Modeling public mood and emotion : Twitter sentiment and socio-economic phenomena. *CoRR, abs/0911.1583*.

**[Bertier et al. 2009]** Marin Bertier, Rachid Guerraoui, Vincent Leroy, and Anne-Marie Kermarrec. Toward personalized query expansion. In Proceedings of the Second ACM EuroSys Workshop on Social Network Systems, SNS '09, pages 7–12, New York, NY, USA, 2009. ACM.

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

**[Damak et al.2014]** Damak, F., Pinel-Sauvagnat, K., Cabanac, G., et Boughanem, M. (2013). Effectiveness of State-of-the-art Features for Microblog Search. In *SAC'13 : ACM Symposium on Applied Computing*. ACM.

**[Damak et al.2011]** Damak, F., Jabeur, L. B., Cabanac, G., Pinel-Sauvagnat, K., Lechani, L., et Boughanem, M. (2011). IRIT at TREC Microblog 2011. In E. M. Voorhees et (Eds.), *Text REtrieval Conference (TREC), Gaithersburg, USA,*. National Institute of Standards and Technology (NIST).

**[Efron, M. 2011b].** The university of illinois graduate school of library and information science at TREC 2011. In *TREC'11 : 20th Text Retrieval Conference.* National Institute of Standards and Technology (NIST).

**[Hammache.2013]** Hammache A: Recherche d'Information: un modèle de langue combinant mots simple et mots composés. Thèse doctorat, UMMTO, 2013.

**[Jansen et al. 2009]** Jansen, B. J., Zhang, M., Sobel, K., et Chowdury, A. (2009). Twitter power : Tweets as electronic word of mouth. *J. Am. Soc. Inf. Sci. Technol.*, *60* (11), 2169–2188.

**[Jabeur et al.2012]** Jabeur, L., Tamine, L., et Boughanem, M. (2012). Featured tweet search : Modeling time and social influence for microblog retrieval. In *IEEE/WIC/ACM International Conference on Web Intelligence, Macau, China* (pp. 166–173). IEEE Computer Society - Conference Publishing Services.

**[Kanhabua, N. et Nørvåg, K. 2011].** A comparison of time-aware ranking methods. In Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '11, pages 1257–1258, New York, NY, USA. ACM.

**[Li, H. 2011].** Learning to Rank for Information Retrieval and Natural Language Processing. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers.

**[Lin et al. 2011]**YuanLin,HongfeiLin,SongJin,andZhengYe. Socialannotationin queryexpansion:amachinelearningapproach.InProceedingsofthe34thinternationalACM SIGIRconferenceonResearchanddevelopmentinInformationRetrieval,SIGIR'11,pages405–414,NewYork,NY,USA, 2011.ACM.

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