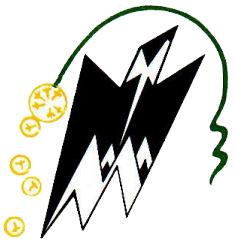
République Algérienne Démocratique Et Populaire
Université Mouloud Mammeri de Tizi-Ouzou
Faculté de Génie Électrique et Informatique
Département d'Informatique



Mémoire de Fin d'Etudes
de MASTER ACADEMIQUE

Domaine : **Mathématiques et Informatique**
Filière : **Informatique**
Spécialité : **Conduite de Projets Informatiques**

Présenté Par
Jugurtha AIT OUFELLA
Lyes ALLACHE

Thème

Utilisation des signaux sociaux pour la recherche d'information dans les microblogs
Cas : Twitter

Président(e) : M^{me}F.Bouarab
Promotrice : M^{me} F.Amirouche
Examinateur : M^{me} L.Belkacemi
Examinateur : Mr M.Amirouche

Table des matières

Table des figures

Introduction générale

Introduction

Introduction

Au cours du vingtième siècle, les individus ont reçu des informations par le biais du bouche-à-oreille, la radio et la télévision ou les éditeurs de journaux et de livres.

Aujourd'hui, avec le développement des technologies et la disponibilité croissante d'Internet, l'information se répand instantanément à travers la planète ;des informations qui étaient difficilement accessibles dans le passé peuvent être aujourd'hui publiées.

Avec l'avènement du web 2.0 et l'apparition des réseaux sociaux et des plateformes de partage, les internautes peuvent dorénavant créer, partager et diffuser des informations.

En effet, les réseaux sociaux sont devenus des espaces d'expression pour des millions de gens sur Internet, ils mettent à disposition de leurs utilisateurs des outils faciles d'utilisation pour que ceux-ci puissent s'exprimer et interagir entre eux.

Au-delà de ces aspects d'utilisation personnelle à des fins de divertissement, les réseaux sociaux offrent aux entreprises et aux communautés virtuelles un moyen de collaboration rapide et pratique. Ils sont maintenant reconnus comme un moyen important pour la diffusion de l'information.

Contexte

Dans notre travail, nous allons nous intéresser à la recherche d'information sociale, plus précisément à la recherche d'informations dans les microblogs.

Les plateformes de microblogging offrent un service en ligne qui permet à leurs utilisateurs de publier des messages de petite longueur.

Parmi les plateformes de microblogging, Twitter est la plus utilisée, elle compte plus de 300 millions d'utilisateurs actifs par mois avec 500 millions de tweets envoyés par jour et est disponible dans plus de 35 langues. Sur Twitter, un utilisateur X peut suivre le flux de tweets envoyés par un utilisateur Y sans avoir besoin de sa permission. Les relations entre utilisateurs des réseaux sociaux sont appelées des abonnements. Si X est abonné à Y, alors X reçoit automatiquement toutes les publications de Y.

Un tweet est étroitement lié à la personne qui le publie, plus un utilisateur a d'abonnées, plus sa publication est susceptible d'être vue et rediffusée ou bien mise en favori. Au sein de la plateforme Twitter, les tweets sont limités à 140 caractères et peuvent contenir des caractères spéciaux tels que :

— Les hashtags qui sont représentés par un '#' suivi par un mot.
— Les RT qui indiquent qu'un message est un retweet.
— Des URL : des hyperliens qui dirigent vers d'autres pages web.

Cette plateforme affiche donc des caractéristiques particulières, les approches de recherche d'information classiques peuvent se trouver obsolètes dans ce contexte, ce qui soulève la problématique lié à la recherche dans les microblogs.

Problématique

Étant donné, l'aspect particulier des microblogs, les méthodes de recherche d'information dites «classiques »ne sont plus capables de restituer l'information de manière efficace, de ce fait, de nouvelles problématiques se posent quant aux méthodes qu'il faut utiliser pour parvenir à des résultats satisfaisants.

— **Quel est le modèle de RI le plus adapté pour gérer les spécificités des microblogs ?** En effet, les modèles de RI dit «Classiques » qui se basent sur les statistiques de fréquence de termes dans les documents et la longueur de ceux-là ne sont pas efficace face à la faible longueur des microblogs. Quelles sont donc les solutions pour y remédier ?
— **Comment évaluer l'influence sociale d'un utilisateur ?** Un utilisateur de Twitter a des followers, d'autres utilisateurs qui se sont abonnés à ses contenus et qui reçoivent les tweets qu'il diffuse, plus un utilisateur a d'abonnées, plus la portée de ses publications est grande. Comment peut-on évaluer l'influence d'un utilisateur au sein d'une communauté ?
— **Comment utiliser les signaux sociaux afin de juger de la pertinence d'une information ?** Un microblog peut être rediffusé, mis en favori par un utilisateur, ou bien partagé sur une autre plateforme, comment peut-on utiliser ces propriétés afin d'améliorer les résultats d'une requête ?

Objectifs

Notre travail vise à améliorer la qualité des résultats de la recherche d'information dans les microblogs.
La recherche consiste en la restitution des microblogs les plus pertinents par rapport à une requête exprimée sous forme de mots-clés.

afin d'atteindre nos objectifs, nous allons effectuer les tâches suivantes :

— Nous allons tout d'abord faire une étude de l'état de l'art de la recherche d'information classique puis de la recherche d'information dans les microblogs.
— En seconde partie de notre travail, nous allons faire une étude des facteurs de pertinence pour l'identification d'un microblog pertinent, nous allons nous intéresser aux signaux sociaux (influence de l'auteur, retweet,la mise en favori) et à leur utilisation afin d'améliorer la restitution de microblogs pertinents.
— Puis nous proposerons un modèle combinant des facteurs de pertinence sociaux ,temporels et thématiques que nous évaluerons par des expérimentations.

Organisation du mémoire

Ce présent manuscrit s'articule autour de deux parties principales contenant deux chapitres chacune.

Dans la première partie intitulée « **De la recherche d'information classique à recherche d'information sociale dans les microblogs** », nous allons définir les concepts de la recherche d'information, puis nous allons dresser un état de l'art de la recherche d'information sociale et plus spécifiquement la recherche d'information dans les microblogs.

La seconde partie intitulée «**Étude des facteurs de pertinence des microblogs et utilisation des signaux sociaux pour l'amélioration des résultats**»quant à elle va détailler notre contribution, à savoir l'utilisation des signaux sociaux dans la recherche dans les microblogs.

Le mémoire sera structuré en cinq chapitres :

Le premier chapitre abordera une vision globale du domaine de la recherche d'informations , ainsi que les concepts fondamentaux à savoir, les étapes principales constituant

un système de recherche d'information, la notion de pertinence et les principaux modèles de la RI connus à ce jour.

Le second chapitre va aborder la recherche d'information sociale, et en particulier la recherche d'information dans les microblogs.

Le troisième chapitre abordera l'état de l'art de la recherche dans les microblogs, nous y décrirons des approches utilisées dans la restitution de microblogs pertinents.

Dans le quatrième chapitre , nous décrirons la solution qu'on a proposé ainsi que l'architecture du modèle de recherche qu'on a utilisé

le cinquième chapitre abordera la phase d'implémentation de notre solution ainsi que son évaluation par le biais d'un ensemble d'expérimentations.

En conclusion, nous dresserons un bilan de notre travail, et introduirons nos perspectives et nos propositions.

Première partie

De la recherche d'information classique à la recherche d'information sociale dans les microblogs

Chapitre 1

Généralités sur la recherche d'information

Introduction

La recherche d'information (RI) est une branche de l'informatique qui s'intéresse à la représentation, l'organisation et l'accès à l'information, elle peut être définie comme une activité dont la finalité est de localiser et de délivrer un ensemble de documents à un utilisateur en fonction de son besoin en informations.

Elle vise à satisfaire les exigences des utilisateurs en informations en mettant en place un mécanisme qui va faire correspondre le besoin de l'utilisateur et les documents d'une base documentaire.

Ce chapitre a pour objectif de présenter brièvement les concepts de base de la RI.

Nous allons d'abord définir ce qu'est un système de recherche d'information ainsi que les différentes notions s'y rattachant, puis nous nous intéresserons au processus général de la RI.

Nous passerons ensuite en revue les principaux modèles de la recherche d'information, enfin nous aborderons les méthodes d'évaluation d'un système de recherche d'information (SRI).

1.1 Système de recherche d'information

Un système de recherche d'information est un ensemble de logiciels qui offrent des techniques et des outils qui permettent de sélectionner, à partir d'un besoin en information d'un utilisateur exprimé sous forme de requête en langage naturel, les documents qui peuvent l'intéresser.

Plus précisément, un système de recherche d'information est un système qui permet

de retrouver automatiquement les documents pertinents à partir d'une grande collection de documents pour une requête donnée.

De cette définition, on dégage quelques concepts clés auxquels s'articule un système de recherche d'information qui sont : le document, la collection de documents, la requête, et la relation de pertinence entre un document et une requête.

- **La collection de documents :** appelée aussi fond documentaire ou corpus, elle constitue l'ensemble des informations ou documents accessibles et exploitables.
- **Document :** le document constitue l'information élémentaire d'une collection de documents, on parle aussi de granule de document qui peut représenter tout ou une partie d'un document. De manière plus générale, c'est toute source d'information susceptible de constituer une réponse à une requête d'un utilisateur.
- **Requête :** Le besoin en information ressenti par l'utilisateur le pousse vers la formulation d'une requête pour interroger la base documentaire. La requête est l'expression du besoin en information de l'utilisateur. Une requête est un ensemble de mots clés, c'est une description sommaire des documents ciblés par la recherche. Elle peut être formulée en langage naturel en spécifiant les mots clés ou bien une expression particulière.

  Dans la RI traditionnelle, le besoin en information de l'utilisateur se compose d'un ensemble de mots-clés, c'est-à-dire que les informations ciblent le contenu textuel d'un document.
- **Pertinence :** Elle peut être définie comme la correspondance entre un document et une requête, ou bien le degré d'utilité du document pour l'utilisateur. La pertinence permet d'évaluer jusqu'à quel point les documents restitués traitent du sujet de la requête.

  Cette notion de pertinence est le pivot de la RI car toutes les évaluations tournent autour d'elle. Elle peut être perçue selon deux niveaux différents à savoir la pertinence système et la pertinence utilisateur.
  - **La pertinence système :**Elle est définie à travers les modèles de RI, elle est souvent traduite par un score évaluant l'adéquation du contenu des documents vis-à-vis de celui de la requête.
  - **La pertinence utilisateur :**Elle est liée à la perception de l'utilisateur sur l'information renvoyée par le système, un document est pertinent pour un utilisateur suivant la satisfaction de son besoin en information.

    Deux utilisateurs peuvent juger différemment un même document, celle-ci est donc subjective.

1.2 Processus de la RI

Pour répondre aux besoins en information de l'utilisateur, un SRI met en œuvre un certain nombre de processus pour réaliser la mise en correspondance des informations contenues dans les documents textuels d'une collection de documents, et les besoins en information des utilisateurs exprimés sous forme de requête.

Un Système de Recherche d'Information (SRI) intègre trois fonctions principales représentées schématiquement par le processus classique en U :

- **L'indexation :** elle a pour rôle d'extraire à partir d'un document ou d'une requête, une représentation paramétrée qui couvre au mieux son contenu.
- **La reformulation de la requête :**Dans un SRI l'utilisateur interroge la collection de documents à travers une requête qui exprime son besoin informationnel. L'utilisateur formule sa requête dans un langage naturel et celle-ci est par la suite représentée sous une forme interne compréhensible par le système.
  Une reformulation de la requête peut parfois être possible, celle-ci consiste à améliorer la requête initiale en rajoutant de nouveaux termes ou en supprimant des termes inutiles, dans but d'améliorer les résultats.
- **L'appariement requête-document :**Consiste en la comparaison entre la représentation du document et la requête, cela revient à mesurer la similarité entre la requête et le document, dans le but de représenter la pertinence du document vis-à-vis de la requête.

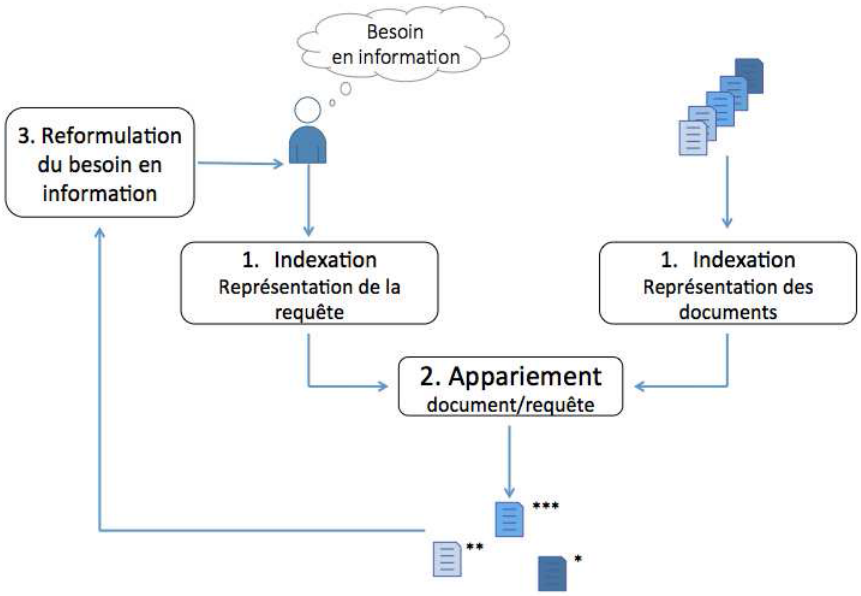


FIGURE 1.1 – Le processus de RI en U

1.2.1 L'indexation

L'indexation consiste à analyser le document ou la requête pour extraire une représentation paramétrée qui couvre au mieux son contenu sémantique.

Cette représentation est souvent une liste de mots clés significatifs, que l'on nomme : descripteurs du document ou de la requête.

Elle consiste donc à représenter le document par un ensemble de mots clés qui résument son contenu d'une manière intelligente, permettant ainsi de le retrouver facilement et rapidement.

La finalité de l'étape de l'indexation est la construction d'un index, ce dernier est un ensemble de descripteurs qui renvoie chacun à des documents. Cet index sera plus facilement exploitable par le système lors du processus ultérieur de recherche. On distingue trois types d'indexation :

- **Manuelle :** Réalisée par un expert documentaliste, cet expert détermine en fonction de ses connaissances, les mots clés qui lui semblent les plus significatifs pour représenter le document.
  Elle garantit une bonne précision cependant, elle présente plusieurs inconvénients liés notamment à l'effort et le prix qu'elle exige.
  L'indexation manuelle est caractérisée par un haut degré de subjectivité lié au facteur humain du fait que pour un même document, des termes différents peuvent être affectés par des indexeurs différents.
- **Automatique :**L'indexation automatique implique une analyse automatique du contenu de chaque document de la collection. Cette analyse comprend plusieurs étapes, le but étant d'extraire les termes représentatifs du contenu et d'évaluer leur pouvoir de représentation du contenu ainsi que leur pouvoir de caractérisation du document dans lequel ils apparaissent.
  Cette approche est la plus communément utilisée.
- **Semi-automatique :**les termes du document sont extraits en un premier temps par un processus automatique, puis le spécialiste du domaine intervient pour effectuer le choix final des termes significatifs et établir les relations entre les mots clés.

L'approche d'indexation la plus utilisée est l'indexation automatique, nous allons nous y intéresser tout particulièrement.

L'indexation automatique effectue un ensemble de traitements sur un document : l'extraction automatique des termes du document, l'élimination des mots vides, la normalisation, la pondération et enfin la création de l'index.

L'extraction automatique des termes

Cette opération consiste à extraire du document un ensemble de termes ou de mots simples par une analyse lexicale permettant d'identifier les termes en reconnaissant les espaces de séparation des mots, des caractères spéciaux, des chiffres, les ponctuations.

L'analyse lexicale transforme donc une suite de caractères en une suite de mots reconnaissables.

L'élimination des mots vides

La liste des mots simples extraite précédemment peut contenir des mots non significatifs, appelés "mots vides", ce sont des mots qui ne traitent pas le sujet d'un document tels que : les pronoms personnels, les prépositions.

L'élimination de ces mots peut se faire soit en utilisant une liste prédéfinie de mots vides (aussi appelée anti-dictionnaire ou stop-list) ou bien en écartant les mots dépassant un certain nombre d'occurrences dans la collection.

Cette étape permet de réduire l'index, on gagne alors en espace mémoire et en temps d'exécution puisque les mots vides ne seront plus traités.

La normalisation (lemmatisation et radicalisation)

On peut trouver dans un texte plusieurs formes différentes d'un mot, toutes relatives à un même sens. La normalisation consiste à représenter ces différentes variantes morphologiques d'un mot par une forme unique (racine grammaticale). Il n'est donc plus nécessaire d'indexer tous ces mots puisqu'un seul suffirait à représenter le concept véhiculé. Ainsi les documents contenant différentes formes d'un même mot auront les mêmes chances d'être restitués.

Il reste cependant à mentionner que ces traitements ont certains inconvénients tels que la perte de sens, du fait que la racine extraite peut être commune à des mots se rapportant à des concepts différents : les mots "port" et "portes" ont la même racine mais expriment des concepts différents.

La pondération

La pondération est une procédure permettant de mesurer l'importance d'un mot dans un document donné pour caractériser son influence sur la représentation de document.

Elle associe une valeur numérique à chaque mot de manière à représenter son importance dans un document de la collection.

Cette importance est souvent calculée sur la base d'aspects statistiques.

Il y a plusieurs approches pour cela :

— **Approches basées sur la fréquence locale (Tf) :**L'objectif de ces approches est de trouver les mots qui représentent le mieux le contenu d'un document. Il est généralement admis qu'un mot qui apparaît souvent dans un texte représente un concept important. il est donc logique de choisir les mots représentatifs selon leurs fréquences d'occurrence, elle est noté $TF_{t,d}$

— **Approches basées sur la fréquence globale (Idf) :** Cette approche mesure la fréquence d'un terme dans toute la collection, c'est-à-dire la pondération globale. En effet, un terme fréquent dans la collection et présent dans beaucoup de documents a moins d'importance qu'un terme moins fréquent, car il ne caractérise aucun document en particulier.
Cette mesure est exprimée selon la formule suivante :

$$IDF_t = log(\frac{N}{n+1}) \tag{1.1}$$

Avec :
N : Le nombre de documents dans la collection.
n : le nombre de documents dans lesquels le terme t apparaît.

— **Approches combinées :** Une nouvelle approche de pondération permet d'avoir une bonne approximation de l'importance d'un terme dans une collection de documents, elle se base sur la combinaison des deux facteurs précédents.
Formellement :

$$TFIDF_{t,d} = TF_{t,d} * IDF_t \tag{1.2}$$

Création de l'index

Au terme du processus d'indexation, un ensemble de structure de données est créé. Ces structures permettent un accès efficace à la représentation des documents.

1.2.2 L'appariement requête-document

La correspondance entre les termes de la requête d'un utilisateur et ceux des documents s'effectue au niveau de l'appariement document-requête.

Une fonction d'appariement détermine le degré de ressemblance d'un document par rapport à une requête, et permet éventuellement de classer les documents par ordre de pertinence pour la requête.

Cette fonction est notée $RSV(q, d)$ (Retrieval Status Value), où q est une requête et d est un document de la collection.

1.2.3 Reformulation de la requête :

Il est parfois possible que l'utilisateur exprime mal sa requête ou qu'il soit incapable de trouver les mots précis pour exprimer son besoin en information. Par conséquent, les résultats que lui fournit la RI ne lui conviennent parfois pas.

Afin de faire correspondre au mieux la pertinence utilisateur et la pertinence système, une étape de reformulation de la requête est souvent utilisée.

La reformulation ou expansion de requêtes consiste à modifier la requête initiale de l'utilisateur par l'ajout de termes significatifs ou la réestimation de leur poids.

Cette étape peut être effectuée :

— Manuellement, dans le cas où l'utilisateur soumet lui-même une nouvelle requête.

— De façon automatique par injection de pertinence :Permet une reformulation de requête initiale, sur la base des jugements de pertinence de l'utilisateur, sur des documents restitués par le système comme réponse à la requête initiale.

Cette méthode consiste en la sélection des termes importants appartenant aux documents jugés pertinents par l'utilisateur, et leur exploitation dans la nouvelle formulation de la requête.

1.3 Modèles de la recherche d'information

Un modèle de RI a pour rôle de fournir une formalisation du processus de RI et un cadre théorique pour la modélisation de la mesure de pertinence.

Un modèle de RI est défini par un quadruplet $(D, Q, F, RSV(q, d))$ où :

— D est l'ensemble de documents.

— Q est l'ensemble de requêtes.

— F est le schéma du modèle théorique de représentation des documents et des requêtes.

— $RSV(q, d)$ est la fonction de pertinence du document d par rapport à la requête q.

Il existe un grand nombre de modèles de RI. Ces modèles ont en commun le vocabulaire d'indexation basé sur le formalisme mots clés et diffèrent principalement par le modèle d'appariement requête-document.

Ils peuvent être classés en trois catégories principales : modèles booléens, modèles vectoriels et modèles probabilistes.

— Les modèles booléens sont inspirés de la logique booléenne et de la théorie des ensembles pour modéliser l'appariement document-requête. Trois variations principales y sont distinguées : le modèle booléen classique, le modèle booléen étendu et le modèle booléen flou.

— Les modèles vectoriels, ils sont basés sur l'algèbre, ils modélisent les documents et les requêtes comme des vecteurs de termes dans un espace multidimensionnel.
Ils englobent le modèle vectoriel généralisé, le modèle LSI (Latent Semantic Indexing) et le modèle connexionniste.

— Les modèles probabilistes sont basés sur la théorie des probabilités, ils ont été introduits pour modéliser la notion de pertinence. Ils englobent le modèle probabiliste général, le modèle de réseau inférentiel (Document Network), et le modèle de langue.

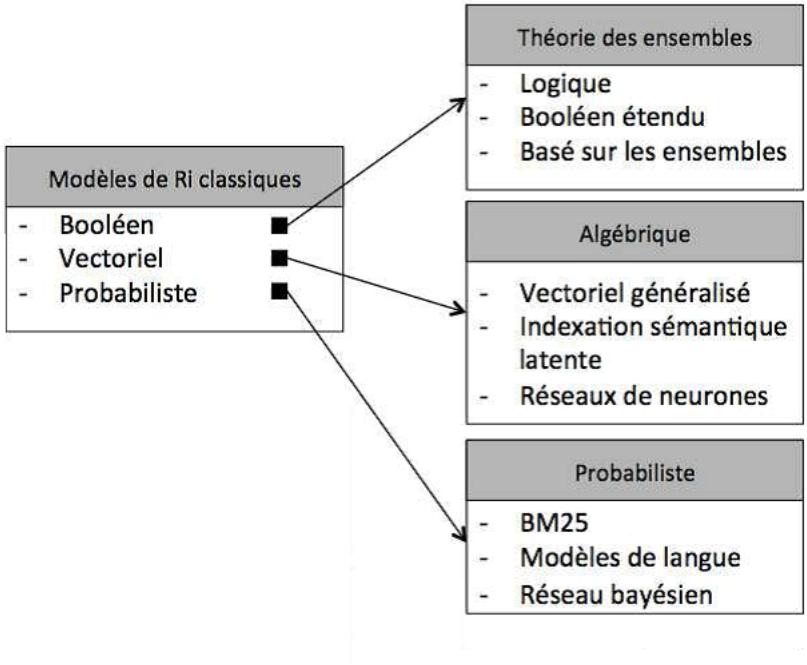


FIGURE 1.2 – Modèles de recherche d'information

1.3.1 Le modèle booléen

Le modèle booléen est le plus simple des modèles de RI, il est basé sur la théorie des ensembles et l'algèbre booléenne.

Chaque document est représenté par une conjonction logique des termes non pondérés qui constitue l'index du document.

Par exemple un document D est représenté comme suit :

$$D = t_1 \wedge t_2 \ldots \wedge t_n$$

Avec t_i Les termes (non pondérés) du document D.

La requête définissant le besoin d'information est une expression logique contenant les termes recherchés qui sont liés entre eux par des opérateurs logiques **(OR, AND, NOT)**.

Un document est alors jugé pertinent et est retourné comme résultat si sa représentation satisfait l'expression logique de la requête, dans le cas échéant, il sera considéré comme non pertinent.

La fonction de similarité entre la requête q et le document d, souvent appelée RSV (Retrieval Status Value), est donnée par la formule suivante :

$$RSV(d, t_i) = \begin{cases} 1 & \text{si } t_i \in d \\ 0 & \text{sinon} \end{cases} \tag{1.3}$$

$$RSV(d, q_i \wedge q_j) = RSV(d, q_i) \wedge RSV(d, q_j)$$

$$RSV(d, q_i \vee q_j) = RSV(d, q_i) \vee RSV(d, q_j)$$

$$RSV(d, \neg q_i) = 1 - RSV(d, q_i)$$

La relation de pertinence dans ce modèle est binaire. Elle est vraie ou fausse. En effet le système retourne un ensemble de documents non ordonnés comme réponse à une requête. Si cette liste est longue, l'utilisateur doit encore fouiller dans cette liste non ordonnée pour identifier les documents qui sont vraiment pertinents à ses yeux.

D'autre part, le nombre d'occurrences d'un terme dans un document n'est pas pris en compte dans ce modèle ($t_i \wedge t_i = t_i$).

Le modèle booléen standard est assez puissant pour des usagers capables de formuler leurs besoins d'information d'une façon concise et précise.

Il présente d'importants avantages tels que simplicité de mise en œuvre et la clarté de l'expression de la requête grâce à des opérateurs logiques, cependant il présente quelques inconvénients qui sont :

— L'absence de classement des documents : en effet l'ordre des résultats n'est pas pris en compte, car les premiers documents retrouvés sont présentés en premier.

— L'expression logique devient compliquée lorsque la requête est longue.

Pour contourner ces lacunes, des extensions sont proposées comme le modèle booléen étendu et le modèle booléen basé sur les ensembles flous[35].

1.3.2 Le modèle vectoriel

Basé sur l'algèbre, il propose une représentation vectorielle pour le document et la requête. La mise en correspondance entre le document et la requête consiste à calculer la similarité entre les vecteurs correspondants.

Formellement, les documents et les requêtes sont représentés dans un même espace,

défini par un ensemble de termes d'index de dimension N comme suit :

$$D =< w_{d1}, w_{d2}, \ldots, w_{dN} >$$

$$Q =< w_{q1}, w_{q2}, \ldots, w_{qN} >$$

Avec :

— w_{di} : le poids du terme t_i dans le document.

— w_{qi} : le poids du terme qi dans le document.

— N :la taille du vocabulaire ou le nombre total de termes de l'index.

Pour déterminer les documents pertinents vis-à-vis d'une requête donnée, le modèle vectoriel repose sur le calcul de la similarité entre le vecteur document et le vecteur requête. Plus les deux vecteurs sont proches, plus la probabilité que le document soit pertinent par rapport à la requête est grande.

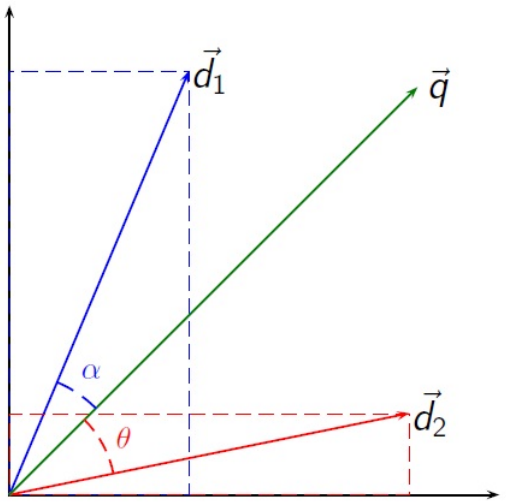


FIGURE 1.3 – Le modèle vectoriel

Une mesure classique utilisée dans le modèle vectoriel afin de mesurer la pertinence est la mesure du cosinus de l'angle formé par les deux vecteurs.

En effet, plus l'angle est aigu, plus le document est proche de la requête, donc, plus la similarité est grande, car plus l'angle formé est petit, et plus le cosinus de cet angle est grand.

Formellement :

$$cos(\vec{d_j}, \vec{q}) = \frac{\vec{d_j}.\vec{q}}{\|\vec{d_j}\|.\|\vec{q}\|} = \frac{\sum_{i=1}^{n} w_{i,j} \times w_{i,q}}{\sqrt{\sum_{i=1}^{n} w_{i,j}^2} \times \sqrt{\sum_{i=1}^{n} w_{i,q}^2}} \tag{1.4}$$

D'autres mesures sont utilisées pour obtenir le score de pertinence d'un vecteur docu-

ment par rapport à un vecteur requête, on peut citer :

— Le produit scalaire qui se calcule comme suit :

$$RSV(q_i, d_j) = \sum_{k=1}^{M} w_{ki}.w_{kj} \tag{1.5}$$

— La mesure de Jacard :

$$RSV(q_i, d_j) = \frac{\sum_{k=1}^{M} w_{ki}.w_{kj}}{\sum_{k=1}^{M} w_{ki}^2 + \sum_{k=1}^{M} w_{kj}^2 - \sum_{k=1}^{M} w_{ki}.w_{kj}} \tag{1.6}$$

Les modèles vectoriels présentent plusieurs avantages par rapport aux modèles booléens cités précédemment, on peut citer :

— Les termes dans le même document sont pondérés selon différents facteurs, ce qui permet de mettre en évidence les sujets les plus importants du document.

— L'utilisation des mesures de corrélation vectorielle permet de trier les résultats selon leur niveau de pertinence par rapport à la requête.

L'inconvénient majeur du modèle vectoriel concerne l'hypothèse d'indépendance entre les termes d'un même document (ou d'une même requête), alors que les termes dans les documents sont souvent sémantiquement liés.

De plus, dans un texte, l'ordre des mots n'est pas pris en compte.

1.3.3 Le modèle probabiliste

Le modèle probabiliste est basé sur la théorie des probabilités, son but est de calculer la probabilité de pertinence d'un document d par rapport à une requête q.

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.

Ainsi on distingue deux classes de documents pour une requête q :

R l'ensemble des documents pertinents et $\bar{R}$ l'ensemble des documents non pertinents.

Pour un document D, on calcule $P(R|D)$ qui est la probabilité que le document D appartienne à l'ensemble des documents pertinents.

Et on calcule $P(\bar{R}|D)$ qui est la probabilité que le document D appartienne à l'ensemble des documents non-pertinents.

Le score d'appariement entre le document d et la requête Q, noté (d, Q) est donné par la formule suivante :

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

En utilisant la formule de Bayes et en simplifiant, on obtient :

$$RSV(d,Q) = \frac{P(d/R)}{P(d/\bar{R})} \tag{1.8}$$

Ou $P(d/R)$ désigne la probabilité de pertinence et $P(d/\bar{R})$ la probabilité de non-pertinence.

Plusieurs solutions ont été proposées pour représenter le document D et pour estimer ces paramètres. La plus connue est celle du modèle BIR (Binary Independance Retrieval) (van Rijsbergen [37]).

Dans cette approche on considère un document d comme une variable représentée par un ensemble d'événements qui dénotent la présence ($x_i = 1$) ou l'absence ($x_i = 0$) d'un terme dans un document.

$$d = (t_1 = x_1, t_2 = x_2, \ldots, t_n = x_n)$$

— t_i est un terme
— x_i peut être égal à 0 ou à 1 il décrit la présence ou l'absence du terme ti dans le document.

En supposant que ces événements sont indépendants, les probabilités de pertinence et de non-pertinence $P(d/R)$ (resp. $P(d/\bar{R})$) sont calculées comme suit :

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

$$P(d/\bar{R}) = \prod_{i=1}^{i=n} P(t_i = x/\bar{R}) \tag{1.10}$$

La fonction $RSV(d,Q)$ peut s'écrire, après transformation, comme suit :

$$RSV(d,Q) = \sum_{t_i \in T} x_i.log(\frac{p_i(1-q_i)}{q_i(1-p_i)}) \tag{1.11}$$

où $p_i = P(t_i \in D|R)$ et $q_i = P(t_i \in D|\bar{R})$ toutes deux sont des probabilités estimées en utilisant la distribution de probabilités de pertinence des termes dans un corpus de test [1].

Un des inconvénients de ce modèle est l'impossibilité d'estimer les probabilités si des collections de test ne sont pas disponibles.

Pour y remédier, de nombreux modèles ont vu le jour, notamment le modèle Okapi

1. voir Evaluation des SRI

BM25 (Robertson[31]) dans lequel le calcul du poids d'un terme dans un document intègre des aspects relatifs à la fréquence locale des termes et la longueur des documents.

La formule est la suivante :

$$RSV(d_i, Q) = \sum_{t_j \in Q} tf_j.log(\frac{N - df_j + 0.5}{df_j + 0.5}).\frac{(k_1 + 1)}{k_1.(1 - b + \frac{b.l_i}{avgdl}) + tf_{ij}} \tag{1.12}$$

Avec :

- tf_j est la fréquence d'un terme t_j dans la requête Q.
- tf_{ij} est la fréquence de tj dans le document d_i.
- df_j est le nombre de documents contenant le terme t_j.
- N est le nombre de documents dans le corpus.
- l_i est la longueur du document di (le nombre de termes d'indexation).
- $Avgdl$ est la moyenne des longueurs des documents dans le corpus.
- k_1 et b sont deux constantes qui dépendent de la collection ainsi que du type des requêtes.

1.4 Évalutation des SRI

L'évaluation d'un SRI se mesure indépendamment de la méthode d'indexation ou du modèle qu'il l'implante. Pour cela, les techniques d'évaluation s'appuient essentiellement sur l'estimation de la qualité des informations retrouvées par le SRI.

La qualité d'un système doit être mesurée en comparant les réponses du système avec les réponses idéales que l'utilisateur espère recevoir. Plus les réponses du système correspondent à celles que l'utilisateur espère, plus ce système est performant.

L'évaluation constitue une étape importante lors de la mise en œuvre d'un modèle de recherche d'information puisque celle-ci permet de paramétrer le modèle et d'estimer l'impact de chacune de ses caractéristiques et enfin de fournir des éléments de comparaison entre modèles.

1.4.1 Mesures d'évaluation

Pour tout système informatique, les mesures les plus courantes de performance sont le temps et l'espace, les systèmes les plus rapides en temps de réponse et les moins gourmands en espace mémoire sont les plus performants.

Dans le cadre des SRI, on s'intéresse plutôt aux résultats retournés par ce dernier, sans négliger les deux premiers critères de performance qui s'applique à n'importe quel système informatique.

Il existe plusieurs mesures d'évaluation, les deux princpiaux groupes de mesures permettant d'évaluer un SRI sont les mesures non-ordonnées et les mesures ordonnées.

a-Mesures non-ordonnées

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

la précision

La précision mesure la capacité du système à rejeter tous les documents non pertinents. Elle donne une indication sur la proportion des documents pertinents renvoyés par le SRI.

La précision est un indicateur pour la qualité des résultats de la recherche, la valeur de la précision est comprise entre 0 et 1. Elle se calcule alors par la formule suivante :

$$Precision = \frac{| \text{ Documents pertinents restitués } |}{| \text{ Documents restitués } |} \tag{1.13}$$

Le rappel

Le rappel mesure la capacité du SRI à sélectionner tous les documents pertinents. Il donne une indication sur le nombre de documents pertinents trouvés par rapport au nombre total de documents pertinents pour la requête.

La valeur de rappel est comprise entre 0 et 1, et plus le rappel est proche de 1, meilleure est la réponse du SRI.

Sa valeur se calcule par la formule suivante :

$$Rappel = \frac{| \text{ Documents pertinents restitués } |}{| \text{ Documents pertinents } |} \tag{1.14}$$

F-Score (ou F-mesure)

C'est une mesure qui combine la précision et le rappel, nommée F-mesure ou F-score introduite dans (Rijsbergen, 1979[37]) et définie par :

$$F - score = \frac{precision.rappel}{precision + rappel} \tag{1.15}$$

Mesures ordonnées

Ce groupe de mesures prend en compte de l'ordre des résultats, ainsi les mesures sont affectées par l'ordre des documents retournés, parmi ces mesures, La précision@X , La

précision moyenne, la R-précision.

— précision@X : c'est la précision à différents niveaux de coupe. Cette précision mesure la proportion des documents pertinents retrouvés parmi les X premiers documents retournés par le système.

$$P@X(q) = P_t/X \tag{1.16}$$

où P_t est le nombre de documents pertinents.

— R-précision : cette précision mesure la proportion des documents pertinents retrouvés après que R documents ont été retrouvés, où R est le nombre de documents pertinents pour la requête considérée.

— La précision moyenne (Average precision-AP) :c'est la moyenne des valeurs de précisions après chaque document pertinent, elle se calcule comme suit :

$$AP_q = \frac{1}{R}\sum_{i=1}^{N} p(i) \times R(i) \tag{1.17}$$

Où $R(i) = 1$ si le i^{ème} document restitué est pertinent, $R(i) = 0$ si le i^{ème} document restitué est non pertinent, $p(i)$ la précision à i documents restitués. R le nombre de documents pertinents pour la requête q et N le nombre de documents restitué par le système.

— 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.

$$MAP = \frac{\sum_{q \in Q} AP_q}{\mid Q \mid} \tag{1.18}$$

1.4.2 Collections de test

Une collection de test (ou collection de référence) est constituée d'une collection de documents donnée, d'un ensemble de requêtes-tests avec leurs «réponses idéales »associées. Ces réponses idéales appelées aussi jugements de pertinence vont permettre d'évaluer la qualité des réponses fournies par les systèmes à évaluer.

Pour évaluer un système de recherche d'information, il suffira alors de lui soumettre les requêtes-tests, et de comparer les réponses qu'il fournit aux réponses types. En mesurant l'écart entre la réponse du système et la réponse type, on obtiendra une mesure de qualité sur les performances du système de recherche d'information.

Une des collections les plus utilisées en RI est la collection TREC.

TREC

Le projet TREC est un programme international initié au début des années 90 par le NIST (National Institute of Standards and Technology) et le DARPA (Defense Advanced Reserach Projet Agency). Ce programme offre des moyens homogènes d'évaluation des systèmes de recherche d'information (tels que des collections de test, des mesures d'évaluation, des protocoles d'évaluation..).

Son objectif est de proposer un standard pour comparer les différents modèles de RI, indépendamment de la méthode de l'indexation ou bien du modèle qu'ils implémentent,afin de mesurer l'efficacité des SRI de manière standard.

La campagne d'évaluation TREC propose plusieurs tâches ou plusieurs chercheurs peuvent participer et proposer leurs SRI, ces derniers seront jugé selon leur performances vis à vis de la collection.

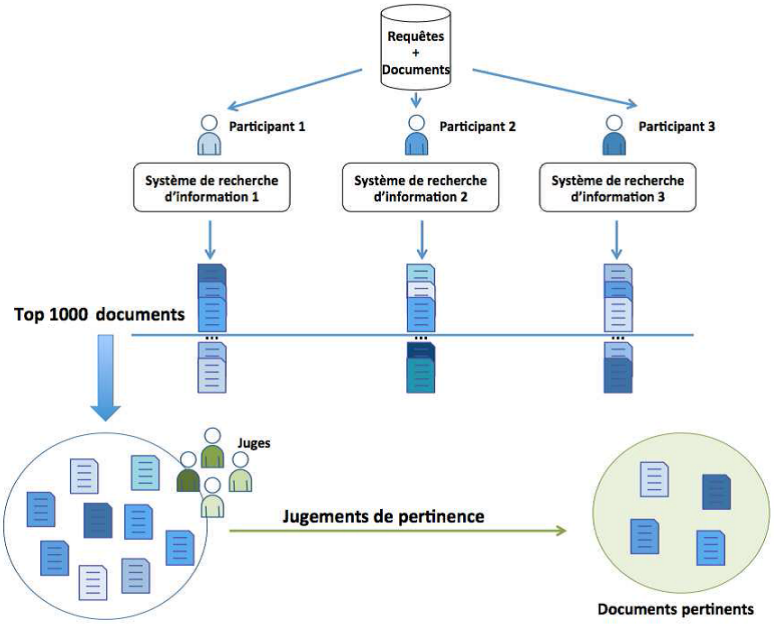


FIGURE 1.4 – Campagne d'évaluation des SRI

Conclusion

Ce premier chapitre a essentiellement porté sur la recherche d'information classique. Nous y avons défini ce qu'est un système de recherche d'information puis nous avons décrit l'architecture commune à tout SRI , puis nous avons brièvement abordé les différentes étapes d'une bonne indexation.

Dans la deuxième partie du chapitre, nous avons vu les différents modèles de RI, ainsi que les différentes méthodes d'évaluation de la performance d'un SRI.

Il en ressort que chacun de ces modèles ou stratégies contribue en partie à la résolution des problèmes inhérents à la recherche d'information : perception du besoin en information, représentation du sens véhiculé par les documents, formalisation de la pertinence.

Dans le prochain chapitre nous allons nous intéresser à une branche spécifique de la RI qui est la thématique principale abordée dans ce mémoire, à savoir, la recherche d'information sociale et plus particulièrement la recherche d'information dans les microblogs.

Chapitre 2

RI sociale et RI dans les microblogs

Introduction

Avec l'émergence du web 2.0 et la démocratisation d'internet, les internautes sont passés de consommateurs à producteurs d'informations. En effet, avec l'apparition des plateformes de partage et des réseaux sociaux, l'utilisateur peut s'exprimer librement sur n'importe quel sujet.

Une grande variété d'applications et de services incitent l'utilisateur à interagir de plus en plus avec les ressources web, de ce fait, une énorme quantité d'informations est générée.

Ainsi avec le volume grandissant de ces données aux spécificités particulières, de nouvelles approches de la recherche d'information sont nées, et celles qui sont les plus adaptées aux contenus sociaux sont regroupées sous l'appellation de "recherche d'information sociale".

2.1 Recherche d'information sociale

La recherche d'information sociale (RIS) est une nouvelle branche de la RI qui est apparue ces dernières années. Elle a pour rôle de retrouver des documents qui correspondent à un besoin d'information d'un utilisateur, tout en intégrant des informations provenant de la participation des utilisateurs des réseaux sociaux dans le processus de recherche.

La RIS a comme objectifs d'améliorer le processus de RI en exploitant les informations sociales et de personnaliser la recherche de l'utilisateur selon son contexte social.

2.1.1 Les informations sociales sur internet

Elles regroupent toutes les informations qui sont soit générées par les utilisateurs eux-mêmes, soit extraites à partir de leur comportement sur internet (les sites qu'ils visitent).

a)Le contenu généré par les utilisateurs (User Generated content)

Ceci définit tout type de contenu tel que les vidéos, les blogs, les discussions de forum, les fichiers audio, ou toute autre forme de média qui a été produit par un internaute en utilisant un outil de production collaboratif.

Parmi ces outils de production de contenu on peut citer :

Blog

Le blog est un site web personnel dans lequel un ou plusieurs auteurs publient au fil du temps des articles (aussi appelés posts ou billets), organisés en catégories et affichés dans l'ordre chronologique inverse. Les visiteurs du blog peuvent ensuite commenter le contenu des articles.

Parmi les plateformes de blogging les plus célèbres : Skyrock, Blogger.

Forum

Un forum est un espace de discussion public ouvert à plusieurs participants.C'est un lieu d'échanges d'informations ou différents utilisateurs apportent leurs contributions en posant des questions ou en répondant à une question posée.Ces contributions forment un fil de discussion.

Ces fils de discussions sont archivés ainsi les utilisateurs peuvent communiquer en différé de manière asynchrone, ils sont généralement classés par thème.

Wiki

Un wiki est un site web dynamique dont tout visiteur peut modifier les pages à volonté. Il permet donc de communiquer ses idées rapidement.

Le principe est simple : il s'agit d'un modèle coopératif de rédaction de documents. Concrètement, n'importe quel visiteur a la possibilité de modifier la page qu'il est en train de lire. Les modifications sont ensuite enregistrées et toutes les versions historiques restent accessibles.

Un premier auteur rédige un article, un second le complète puis un visiteur corrige d'éventuelles erreurs qu'il aura remarquées en navigant sur le site.

Le wiki le plus connu de nos jours est Wikipédia.

Réseaux sociaux numériques

Un réseau social numérique est un site internet qui permet aux internautes de se créer une page personnelle afin de partager et d'échanger des informations, des photos ou des vidéos avec leur communauté d'amis et leur réseau de connaissances. Ces réseaux réunissent des personnes via des services d'échanges personnalisés, chacun pouvant décider de lire les messages de tel ou tel autre utilisateur.

Facebook, créé en 2004, est le plus connu d'entre eux, et le plus utilisé à ce jour.

Microblog

Les microblogs sont des publications sous forme messages assez courts qui sont postés sur des plateformes sociales spécifiques. Parmi les plateformes de microblogging les plus célèbres : Twitter.

b)Contenu généré par la pratique

Ce type d'information est produit des différentes activités de l'utilisateur lors de sa navigation sur internet. Ces informations peuvent être regroupées en :

— Les traces de navigation : ce sont des données implicites dérivées des activités de l'utilisateur lors des sessions de navigation, au travers des pages visitées ou des requêtes soumises.

Ces traces sont généralement des cookies, des petits fichiers texte stockés sur le terminal de l'internaute et qui permettent conserver des données de l'utilisateur.

— Les données personnelles : ce sont des données explicites collectées directement de l'utilisateur au travers de formulaires ou de questionnaires.

2.2 Exploitation des informations sociales sur le web

Il existe deux types d'approches d'exploitation des informations sociales sur le web, les approches qui se basent sur l'étude du réseau social (La RI sociale), et les approches qui exploitent les données sociales pour enrichir les documents et les requêtes.

2.2.1 Processus de la RI sociale

La recherche d'information sociale diffère des autres approches de recherche d'information du fait qu'elle intègre les propriétés et la structure du réseau social dans le processus de recherche d'information.

Ainsi le processus de recherche se déroule en trois grandes étapes : (1)l'extraction de la structure du réseau social, (2)l'analyse du réseau social, et enfin (3)le classement des documents par pertinence.

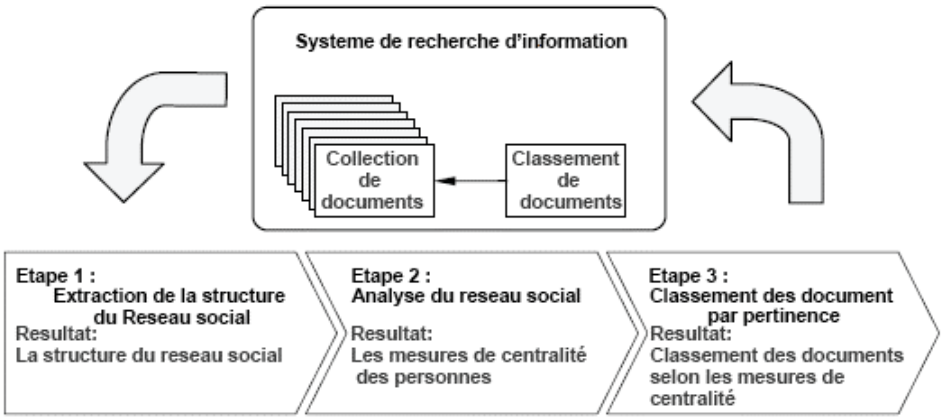


FIGURE 2.1 – Récapitulatif du processus de RIS [34]

1)l'extraction de la structure du réseau social

Dans cette étape, la structure du réseau social est extraite depuis la collection de documents grâce aux relations sociales mises en évidence dans le graphe de contenu social.

Un graphe social sur internet est un modèle ou une représentation d'un réseau social. Il contient les données sociales (documents, commentaires, annotations et votes) ainsi que les interactions mutuelles entre les personnes et ce contenu.

La topologie d'un graphe social peut différer d'un réseau social à un autre, mais de manière générale elle définit deux types d'entités : les acteurs et les données.

Les acteurs représentent les personnes, ils peuvent avoir différents rôles dans le graphe social, et les données représentent généralement des informations.

2)Analyse du réseau social

L'analyse des réseaux sociaux est l'étude des relations sociales entre des individus au sein d'un réseau social.

Elle se focalise sur le comportement d'un acteur plutôt que sur ses attributs ou ses qualités, ceci dû au fait que même si les attributs d'un acteur le décrivent au niveau personnel, ils ne peuvent fournir des informations quant à ses interactions sociales avec les autres utilisateurs. (Wasserman and Faust[38], 1994). De plus des acteurs avec des attributs similaires peuvent avoir un comportement différent sous l'influence de leur cercle social.

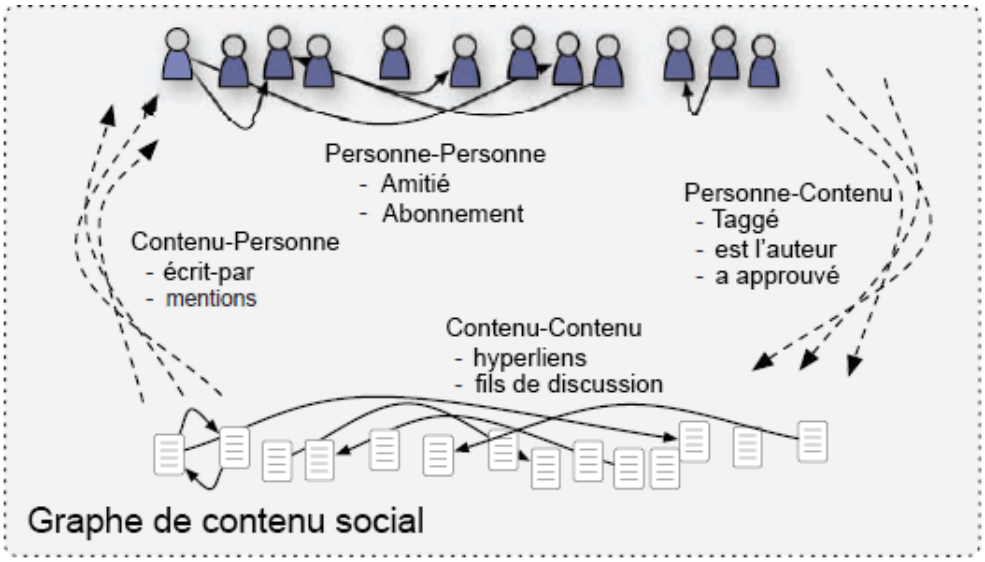


FIGURE 2.2 – Graphe de contenu social

Dans le but d'identifier les acteurs clés dans un réseau social, autrement dit les personnes les plus populaires ou ayant le plus d'influence, un ensemble de mesures dites de centralité a été défini.

Ces mesures décrivent les attributs structurels des nœuds au sein des réseaux sociaux. Elles permettent d'évaluer l'importance d'un acteur par rapport à sa position dans le réseau social. Parmi ces mesures :

— **Le degré de centralité Nieminen (1974) :**
Il est défini comme le nombre de liens incidents à un nœud. Au sein d'un réseau social, cela nous permet de calculer combien une personne a d'abonnés ou d'amis. Il est calculé selon la formule suivante :

$$C_D(v_i) = \sum_{\forall v_j \in V} a(v_i, v_j) \tag{2.1}$$

Avec

$$a(v_i, v_j) = \begin{cases} 1 & \text{s'il existe un arc reliant } v_i \text{ et } v_j \\ 0 & \text{sinon} \end{cases} \tag{2.2}$$

— **La centralité de proximité (Sabidussi, 1966) :**
Elle évalue la proximité de chaque acteur au reste des nœuds du réseau, la centralité de proximité prend en considération les connexions directes ainsi que les connexions indirectes à travers les nœuds intermédiaires.

Elle se calcule par la formule suivante :

$$C(x) = \frac{1}{\sum_y d(v_i, v_j)} \tag{2.3}$$

Avec $d(v_i, v_j)$ est la longueur du plus court chemin entre les nœuds v_i et v_j.

— **La centralité d'intermédiarité (Anthonisse, 1971 ; Freeman, 1977) :**
Elle permet d'évaluer le rôle d'un acteur dans les flux de communication entre les nœuds du réseau social. Elle évalue l'habilité d'un acteur à contrôler un flux de communication, elle permet d'identifier les intermédiaires au sein du réseau. Elle se calcule selon la formule :

$$C_B(v_i) = \sum_{\forall v_j \in V} \sum_{\forall v_k \in V} b_{jk}(v_i) \tag{2.4}$$

Ou $b_{jk}(v_i)$ est la longueur du plus court chemin qui connecte v_j à v_k en passant par v_i.

3)Classement des documents par pertinence

Un score de pertinence par rapport à la requête et un score de pertinence sociale sont calculés, puis combinés afin de produire le classement final des documents comme suit :

$$score(q, d) = Comb(RSV(q, d), S_d) \tag{2.5}$$

Avec $RSV(q, d)$: le score de pertinence thématique
et S_d :le score de pertinence sociale.

2.2.2 Utilisation des informations sociales pour amélioration des résultats

Les informations sociales peuvent être exploitées après les processus de RI classique (classement des documents selon un score de pertinence social) ou pendant même le processus de RI.

Plusieurs solutions ont été proposées afin d'améliorer la qualité des résultats en utilisant les informations sociales. Celle-ci sont classées selon le niveau d'approche qu'elles utilisent à savoir, les approches orientées utilisateurs et les approches orientées document.

— **Approches orientées utilisateurs :** Elles ont pour but d'améliorer les résultats de recherche en exploitant les informations sociales de l'utilisateur afin de mieux cerner son besoin en information.

Certaines approches ont exploité ce contexte social dans la reformulation de requêtes, tandis que d'autres se basent sur la création d'un profil utilisateur pour une recherche personnalisée.

— **Reformulation de requêtes en utilisant les informations sociales :** La reformulation de requête consiste à créer une nouvelle requête plus adéquate que celle initialement formulée par l'utilisateur, une requête étendue permet de mieux exprimer le besoin en information de l'utilisateur, de ce fait, les chances de retourner des documents pertinents sont plus grandes.

Une des méthodes utilisées consiste à s'appuyer sur les documents restitués par le SRI et à compléter la requête par une sélection des termes les plus fréquemment rencontrés dans ces textes.

— **Création d'un profil utilisateur pour la recherche personnalisée :**

L'objectif est d'adapter les processus de recherche et la restitution de documents aux caractéristiques de l'utilisateur, telles que ses préférences et intérêts.

Pour modéliser un profil utilisateur, plusieurs catégories de données peuvent être exploitées.

— Des données explicites collectées directement de l'utilisateur au travers de formulaires par exemple.

— Des données extraites du réseau social dont il est membre, dans ce cas toutes les données du profil existant sont utilisées, à savoir, ses centres d'intérêt, ses informations personnelles ainsi que ses liens sociaux.

— Des données implicites dérivées des activités de l'utilisateur au travers des pages visitées ou des requêtes soumises.

Ces profils permettent de mieux définir les besoins en information de l'utilisateur, plus particulièrement lorsqu'ils sont ambigus, la prise en compte des intérêts de l'utilisateur ayant formulé ce besoin en information peut permettre d'identifier quel aspect de la requête l'intéresse.

— **Approches orientées documents :** Ces approches se basent sur l'utilisation des informations sociales telles que les tags et les commentaires afin d'enrichir le contenu et la représentation des documents d'une part, et d'autre part, ces informations sociales sont utilisées afin d'évaluer la pertinence d'un document et de le classer en adéquation avec sa pertinence sociale.

Conclusion

Dans la première partie de ce chapitre, nous avons abordé la RI sociale et ses différents processus, nous avons aussi décrit quelques approches qui exploitent les informations sociales pour améliorer la restitution de documents pertinents.

Dans la seconde partie de ce chapitre, nous allons nous intéresser à la RI dans un média social particulier, Twitter.

2.3 La recherche d'information dans les microblogs :- Cas de twitter

Les microblogs sont des publications sous forme de messages courts qui sont postés sur des plateformes sociales spécifiques. Ils sont de plus en plus répandus sur internet, les utilisateurs y portent de plus en plus d'intérêt. Les plateformes de microblogging se caractérisent par la diversité d'informations qui y circulent ainsi que l'intensité des interactions sociales entre les individus. Au sein de ces plateformes, un utilisateur peut exprimer son opinion sur un sujet quelconque, ou bien commenter des publications qui ont été partagées par d'autres utilisateurs, il peut aussi partager des photos ou bien des hyperliens. Généralement ces informations ne sont pas instantanément indexées, de ce fait, elles ne sont pas disponibles pour les moteurs de recherche "classiques". En effet, la recherche dans les microblogs est différente de la recherche documentaire sur le web, du fait que les données ont une structure et un format différent. De plus les motivations de recherche sont spécifiques aux microblogs, les requêtes sont souvent motivées par l'activité sociale de la personne concernée ainsi que les tendances et les évènements courants. La pertinence d'une information dépend du contexte social du celle-ci, ainsi une information a quasiment la même importance que la personne qui la publie.
Il existe plusieurs plateformes de microblogging, les plus populaires auprès des internautes sont Tumblr et Twitter.

La plateforme Twitter étant l'objet de notre étude nous nous intéresserons à cette dernière.

2.3.1 Plateforme de microblogging : Twitter

Twitter est l'un des médias sociaux les plus populaires. Il fut lancé en juillet 2006 aux États-Unis. Ses créateurs le définissent ainsi : « Twitter offre à chacun l'opportunité de créer et de partager instantanément des idées et des informations, sans aucune barrière ».

Les utilisateurs inscrits publient des messages limités à 140 caractères, appelés «Tweets», peuvent se lier entre eux selon le principe d'abonnement.

Twitter occupe une place importante dans notre environnement médiatique. Il est de fait devenu un outil de communication prisé de beaucoup de journalistes, acteurs de la vie politique ou encore des entreprises. L'étude menée par des spécialistes en 2009 révèle par exemple le rôle important de Twitter au sein de la stratégie de communication adoptée par Barack Obama durant la campagne présidentielle de 2008 aux États-Unis.

Twitter compte environ 645 millions d'utilisateurs inscrits, dont plus de 300 millions actifs [1] qui publient plus de 60 millions de tweets quotidiennement.

Les utilisateurs ont la possibilité de rédiger et de publier des tweets en temps réel, notamment grâce aux terminaux mobiles.

Les utilisateurs utilisent aussi la plateforme pour chercher des informations sur un sujet particulier ainsi le nombre de requêtes soumises atteint les 2.1 milliards par jour.

Concepts et fonctionnement de Twitter

Pour qu'un internaute crée son compte sur Twitter il se doit d'ajouter ses informations personnelles telles que sa photo, sa localisation, son site Web et une courte bibliographie, ainsi son profil est créé avec les informations qu'il a rentrées.

Une fois l'inscription effectuée il pourra commencer à publier ses messages, en d'autres termes, tweeter.

Chaque tweet apparaît sur la page de profil de son auteur et est instantanément transmis à ses abonnés, qui le reçoivent dans leur « timeline ». La timeline consiste en l'empilement en ordre chronologique inverse des tweets publiés par les utilisateurs suivis.

Les figures suivantes montrent l'exemple d'un profil au sein de Twitter :

1. Twitter considère un utilisateur comme actif à partir du moment où il suit 30 comptes et est suivi par un tiers d'entre eux

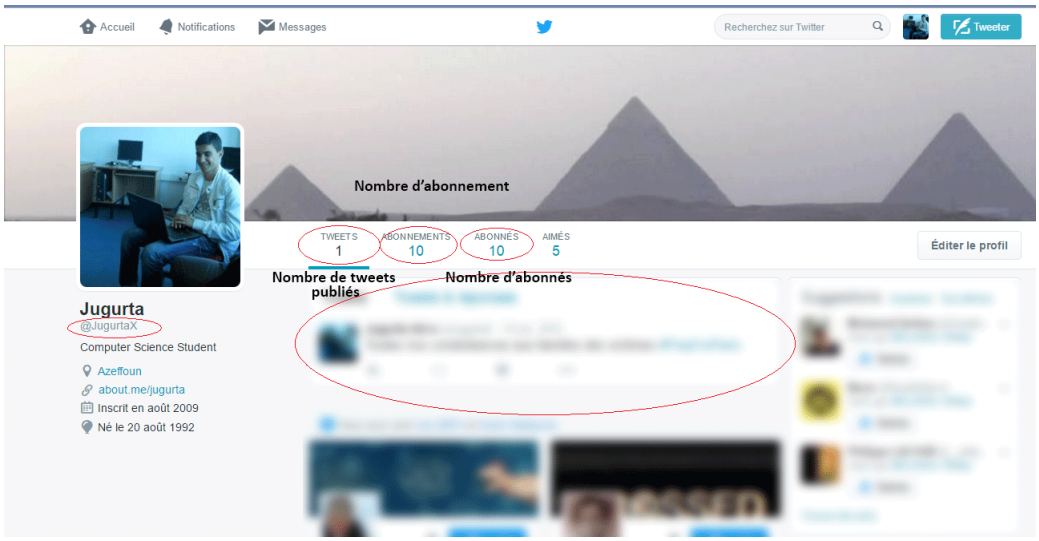


FIGURE 2.3 – Interface de Twitter

Les messages publiés par les utilisateurs de Twitter abordent des thématiques diverses et variées, qu'elles soient d'ordre public – c'est-à-dire en lien avec des éléments d'information susceptibles d'intéresser un large public – ou bien d'ordre personnel.

Les relations au sein de Twitter sont sous forme d'abonnements, ainsi un utilisateur de Twitter X peut suivre le flux de tweets envoyés par un utilisateur Y sans avoir besoin de sa permission. Si X est abonné à Y, alors X reçoit automatiquement toutes les publications de Y.

Chaque tweet est associé à un utilisateur qui représente l'auteur de ce message, néanmoins ce tweet peut être partagé (retweet) par un autre utilisateur qui le juge intéressant.

Caractéristiques de la plateforme Twitter

Accessibilité des données : Twitter, contrairement à la majorité des médias sociaux permet la collecte de données, ainsi chaque internaute peut consulter les messages publiés par un utilisateur de la plateforme sans que celui-ci ait besoin de s'inscrire.

Usage temps-réel : Les utilisateurs de Twitter, ont tendance à publier des informations sur leur quotidien assez régulièrement, il leur arrive de s'exprimer sur des évènements qui viennent de se produire.

De plus, un utilisateur qui est abonné à un autre reçoit les tweets de celui-ci dans sa timeline, au moment même où ceux-ci sont publiés.

Hétérogénéité du contenu : Au sein de la plateforme Twitter, tous les tweets sont limités à 140 caractères.Un utilisateur peut inclure différents symboles dans un tweet,

en plus du contenu textuel qu'il rédige.

Ainsi de nombreux usages ont fait leur apparition :

— @ suivi par un nom d'utilisateur permet d'indiquer qu'on mentionne ou s'adresse à une personne ou entité particulière.

— # suivi par un mot est un hashtag. Un hashtag indique un mot important permet de catégoriser les tweets selon un contexte spécifique.

— Les tweets peuvent également contenir des URL qui renvoient vers des pages web.

— RT indique que le message est un retweet : un utilisateur partage le microblog d'un autre utilisateur s'il le juge intéressant.

— La mise en favori, un utilisateur met en favoris un tweet pour montrer son intérêt au microblog.

— Les utilisateurs peuvent aussi ajouter du contenu multimédia tel que des photos.

Les microblogs contiennent aussi des données qui ne sont pas saisies par l'utilisateur directement telles que :

— Le nom de l'auteur qui a publié le message.

— L'heure à laquelle le microblog a été publié.

— La géolocalisation de l'auteur si celui-ci a activé cette option.

2.3.2 Recherche d'information dans les microblogs

Les moteurs de recherche des microblogs, sont différents des moteurs de recherche du web, ainsi les données en entrée ou en sortie ne sont pas les mêmes. Un utilisateur peut rechercher un hashtag, une personne, des mots clés, une URL ou bien tout mélanger. Les résultats en sortie seront distincts, ainsi il y aura une liste de résultats contenant les microblogs, une liste contenant les utilisateurs et une liste contenant les hashtags. De plus, les motivations des utilisateurs pour la recherche de microblog ne sont pas les mêmes par rapport à la recherche documentaire. Ainis selon une étude effectuée par Jaime Teevan en 2011, où on a observé les habitudes de 54 utilisateurs de Twitter. On a constaté que la moitié des utilisateurs recherchent des informations sur les actualités, tels que les résultats d'une élection, un attentat, un grand accident. Le quart des participants ont recherché des informations sociales, un utilisateur en particulier, ou bien un groupe d'utilisateurs ayant les mêmes intérêts. Tandis que le reste des utilisateurs effectuait des recherches similaires aux recherches sur le web sur un sujet spécifique.

a) Motivations de recherche dans les microblogs

Un important volume d'informations sociales est généré du fait de la popularité des microblogs et de leur production quotidienne. Les chercheurs se sont donc intéressés à

la recherche d'information dans les microblogs, ils ont proposé des approches selon la motivation de recherche.

Ces approches sont généralement classées selon le type d'information recherché.

— **Recherche temps réel dans les microblogs**

Rechercher des informations dans les microblogs nous permet de trouver en temps réel des informations fiables sur évènement qui s'est produit récemment. Généralement, pour qu'un moteur de recherche indexe ces informations, il lui faudrait un délai de quelques jours à une semaine. Le facteur le plus important dans la recherche de microblogs en temps réel est la date de publication du document.

Ainsi, l'une des méthodes de restitution des microblogs pertinents consiste à classer temporellement tous les résultats pertinents et restituer ceux qui sont temporellement proches à la date de soumission de la requête.

— **Recherche d'opinion**

La recherche d'opinion a pour objectif de retrouver les documents exprimant des opinions sur le sujet de la requête, ainsi les tweets restitués doivent satisfaire le besoin en information de l'utilisateur et exprimer une opinion claire à propos du sujet en question.

De nombreux travaux se sont intéressés à la recherche d'opinion dans les microblogs, on peut notamment citer Bernard J. Jansen en 2009 qui a étudié l'opinion des utilisateurs d'un produit d'une marque particulière, il en a résulté que la moitié des tweets avaient une opinion positive.

Ainsi grâce aux microblogs, on peut obtenir des opinions et des réactions immédiates sur des produits ou des marques.

— **Détection de tendances**

La détection de tendances a pour but d'identifier les sujets les plus actifs au sein de réseau social.

Plus particulièrement, elle se base sur la recherche d'expressions fréquentes dans le flux des microblogs et jauge l'intérêt du public pendant une période particulière.

Ayant un lien direct avec les évènements se déroulant dans le monde à un moment donné, une tendance correspond à un évènement de grande envergure qui intéresse un grand nombre d'utilisateurs, tel que les élections ou des évènements sportifs.

La détection de tendances, nous permet aussi de surveiller un évènement spécifique dans les flux des microblogs, ainsi on peut déclencher des alertes ou des avertissements, comme dans le cas d'un séisme ou d'une épidémie.

— **Recherche de microbloggers**

Les activités intensives de microblogging, ont engendré des groupes de microbloggers qui ont une position assez importante au sein du réseau social.

Ils jouent un rôle clé dans leur entourage social ainsi qu'au niveau de réseau social lui-même.
Ainsi l'objectif de la recherche de microbloggers est l'identification des utilisateurs les plus populaires ou bien des experts dans domaines particuliers.

b) Critères de pertinence des microblogs

La recherche de microblogs a pour but de restituer les microblogs contenant l'information la plus pertinence par rapport à une requête.

Afin d'évaluer la pertinence d'un microblog, il existe plusieurs critères à prendre en compte, en plus de la pertinence thématique. Parmi ces critères il y en a qui sont liés au réseau social tels que : la popularité de l'auteur, le facteur temporel.

Dans cette section nous présenterons en détail, les principaux facteurs de pertinence utilisés pour la restitution de microblogs.

- **Critère de pertinence thématique**

  Ce critère est le plus important, la fréquence des termes est utilisée dans ce contexte pour estimer la similarité entre le tweet et la requête.
  Mais vu la faible longueur des microblogs, les modèles de RI classique se retrouvent limités car les termes généralement n'apparaissent pas plus d'une fois dans un microblog.
  Néanmoins certaines approches ont été proposées afin de remédier à ce problème, telles que l'expansion des requêtes ou bien l'enrichissement des microblogs avec des termes de microblogs similaires.

- **Critère de pertinence temporelle**

  Comme évoqué précédemment, la recherche de microblogs est motivée par le besoin d'accès à des informations récentes.
  De ce fait, les tweets les plus récents ont plus de chance d'être pertinents par rapport à des tweets publiés à ceux publiés antérieurement.
  Certaines approches proposent de considérer le temps de soumission de la requête et le temps de publication du tweet, ainsi un tweet publié à la même période que la requête, serait plus pertinent qu'un tweet publié à une autre date.

- **Critère de pertinence sociale**

  Ce critère de pertinence englobe des mesures et des métriques qui décrivent l'importance sociale des utilisateurs et des microblogs.
  De plus la recherche dans les microblogs se doit de considérer la crédibilité de l'information et de ne présenter que des sources fiables.
  Un tweet a plus de chance d'être pertinent et fiable s'il est publié par des utilisateurs importants au sein du réseau social.
  Une approche(Duan et al [9] en 2010) classe les tweets selon le nombre de retweet ainsi que le nombre de mentions tout en prenant en compte la popularité de l'auteur.

- **Critère de pertinence : utilisation de métadonnées**

  Les microblogs contiennent des métadonnées telles que les hashtags (caractérisés par "#") ou bien des mentions à un auteur (à l'aide du caractère "@") ainsi que des URL qui renvoient vers une page web quelconque.
  Certains travaux ont proposé l'utilisation de la présence d'URL ou de métadonnées comme critère de pertinence d'un microblog.
  D'autres approches exploitent ces informations afin d'enrichir la représentation du tweet ou bien dans l'expansion de la requête.

- **Autres critères de pertinence**
  - **La localisation géographique :**
    La localisation géographique de l'endroit ou un tweet a été publié permet d'évaluer sa pertinence. Ainsi un tweet publié à un endroit où se produit un évènement particulier est plus pertinent qu'un tweet publié à plusieurs lieux de là. De plus certains utilisateurs recherchent des tweets autour d'un sujet propre à leur région, ou leur pays, de ce fait, les tweets géographiquement proches de leur localisation ont plus tendance à les intéresser.
  - **Les sentiments :**
    Un tweet qui exprime un sentiment à propos d'une personnalité, d'un produit quelconque peut-être pertinent dans le cas où des utilisateurs s'intéresseraient à l'opinion des gens sur le sujet.
  - **La qualité du tweet :**
    La qualité du tweet prend en considération plusieurs facteurs tels que : la longueur du tweet en calculant nombre de termes dans le microblog, ainsi que la qualité du langage et le vocabulaire utilisé, ceci afin de ne pas restituer les

spams et les tweets ambigus.

Des études ont montré que l'utilisation de ces critères de pertinence lors du processus de la RI ont un impact considérable sur l'efficacité de la recherche, néanmoins, notons qu'un critère de pertinence peut être plus moins important selon le sujet de la requête ainsi que la motivation de recherche.

2.3.3 Evaluation de la RI dans les microblogs

Pour évaluer une approche de RI dans les microblogs, nous aurons recours à une collection de test afin de mettre en œuvre les différentes approches de restitution de microblogs pertinents.

Parmi les compagnes d'évaluation les plus populaires : TREC Microblog.

La tâche TREC Microblog

TREC microblog est une tâche spécifique de la compagne d'évaluation TREC dédiée à la recherche d'information dans les microblogs, qui est organisée annuellement depuis 2011.

Elle évalue les méthodes et les approches de recherche d'information dans les plateformes des microblogging telles que Twitter.

Elle est décrite comme une tache de recherche ad hoc temps réel. Les systèmes doivent donc donner une réponse à la requête en fournissant une liste de documents pertinents classés du plus récent au plus vieux par rapport au moment de soumission de la requête.

Ainsi l'utilisateur va accéder à l'information la plus récente et la plus pertinente vis à vis de sa requête.

La première campagne Trec microblog : TREC 2011, fournit le corpus Tweets2011, qui comprend environ 16 millions de tweets qui ont été publiés sur une période approximative de deux semaines (du 23 Janvier 2011 au 7 Fevrier 2011) . Le corpus est considéré comme un échantillon fiable de la "twittosphère".

Tweets2011 comprend 50 "Topics" (ou requêtes) dont chacun représente un besoin en information à un moment donné. Le format des topics est donné par la figure 2.4.

```
<top>
    <num> Number : MB01 </num>
    <title> Wael Ghonim </title>
    <querytime> 25th February 2011 04 :00 :00 +0000 </querytime>
    <querytweettime> 3857291841983981 </querytweettime>
</top>
```

FIGURE 2.4 – Exemple de topic pour la tâche TREC microblog

où *title* décrit le besoin en information.
querytime représente le moment ou la requête a été soumise.
querytweettime correspond à l'identifiant du dernier tweet soumis avant la requête.

Tweets2011 fournit également un ensemble de jugements de pertinence sur les tweets de la collection. Les tweets sont jugés en fonction du besoin en information spécifié, ils sont classés selon l'ordre suivant :

- Non pertinent : le contenu de ce tweet ne fournit aucune information sur le sujet recherché.
- Pertinent : le tweet contient fournit certaines informations qui peuvent s'avérer utiles.
- Hautement pertinent : ce tweet a une très grande valeur informative sur le sujet recherché.

Conclusion

Dans ce chapitre nous avons présenté les principales approches pour la recherche et la restitution de microblogs, dans un premier temps nous avons abordé la plateforme de microblogging Twitter, et nous avons exploré ses différentes fonctionnalités et nous nous sommes penchés sur ses spécificités.

Puis nous nous sommes interessés aux différentes approches de recherche dans les microblogs, et nous avons exploré les différents facteurs qui font la pertinence d'un microblog.

Enfin nous avons présenté la tâche TREC pour les microblogs.

Dans les prochains chapitres nous allons dresser un état de l'art de RI dans les microblogs.

Chapitre 3

Etat de l'art de la RI dans les microblogs

Introduction

Les usagers des médias sociaux produisent beaucoup d'informations, plus spécialement sur les plateformes de microblogging telles que Twitter, ou pour n'importe quel sujet donné, le nombre de publications atteint facilement la dizaine de milliers. De ce fait, accéder à des informations pertinentes pourrait s'avérer être une tache dure. Ainsi plusieurs chercheurs se sont intéressés à cette problématique et ont essayé d'y apporter leur contribution.

Travaux de recherche d'information dans les microblogs

Généralement, toutes les approches qui tentent d'extraire de l'information des microblogs utilisent des critères différents pour compenser l'impuissance des systèmes de recherche à gérer les nouvelles spécificités de ce média. Ainsi, plusieurs chercheurs ont proposé d'exploiter les critères de pertinence cités plus haut afin d'améliorer et de raffiner les résultats des moteurs usuels de la RI.

Dans ce qui suit nous allons nous intéresser à cinq approches, une approche qui se concentre sur l'exploitation du contenu du tweet et au corpus (Damak [6] 2013), une approche qui exploite le réseau social des microbloggeurs (Ben jabeur [17] 2013), une approche statistique qui a pour but de reclasser les résultats en se basant sur l'exploitation des données statistiques fournies par l'API de twitter (Masaki Aono [1] 2015), une approche de classement par modèle d'apprentissage (Duan et Al [9]) et enfin une approche se basant sur l'expansion des documents et des requêtes(Jamie Callan [18]2012) .

3.1 Approches se basant sur l'exploitation du contenu du tweet

Certaines approches se focalisent sur la structure même du tweet et son contenu, Ainsi, Damak [6] (2013) dans ses travaux a défini cinq familles de facteurs de pertinence pour améliorer les résultats obtenus lors de la recherche :

- **Facteurs de pertinence basés sur le contenu des tweets :**
  - Popularité du tweet (Duan et al.[9], 2010) :
    Ce facteur de pertinence estime la popularité d'un tweet dans un corpus. Un tweet est supposé populaire si plusieurs autres tweets ont un contenu similaire.
  - Longueur du tweet (Duan et al.[9], 2010) :
    Un message plus long est susceptible de véhiculer plus d'information, ce facteur de pertinence est calculé en comptant le nombre de termes dans un tweet.
- **Facteurs de pertinence basés sur l'hyper-textualité :**
  - Présence d'URL dans le tweet (Nagmoti et al.[29], 2010 ; Zhao et al.[14], 2011) :
    Ce facteur suppose que la présence d'une URL indique que le tweet a un caractère informatif renforcé
  - Fréquence de l'URL dans le corpus :
    Ce facteur de pertinence permet de calculer la popularité des URLs publiées dans un tweet ti dans le corpus.
- **Facteurs de pertinence basés sur les hashtags :**
  - Présence d'un hashtag (Metzler et Cai[27] 2011) :
    Un hashtag permet de marquer un contenu avec un mot-clé, ainsi sa présence peut être considérée comme un facteur de pertinence.
  - Fréquence du hashtag du tweet (Duan et al[9], 2010) :
    Ce facteur permet de calculer la popularité du hashtag dans le corpus.
- **Facteurs de pertinence basés sur la popularité des auteurs :** Afin de tenir compte de la popularité des auteurs, deux facteurs ont été pris en compte
  - Nombre de tweets de l'auteur (Nagmoti et al.[29], 2010) :
    ce facteur de pertinence a pour but de valoriser les tweets publiés par des auteurs actifs par rapport aux tweets publiés par des auteurs moins actifs.
  - Nombre de citations de l'auteur (Zhao et al.[14], 2011) : Plus un auteur est mentionné, plus il est populaire, ce facteur évalue la popularité de l'auteur en fonction du nombre de publications ou il est mentionné.
- **Facteurs de pertinence relatifs à la qualité des tweets :**
  - Retweet (Metzler et Cai,[27] 2011) :
    Si un tweet est aimé par un utilisateur, celui le partage à son tour (retweet) le

nouveau message ou bien le message diffusé va être précédé par RT.

— Fraîcheur (Magnani et al.[26], 2012) : Ce facteur se base sur la différence temporelle entre la date de publication du tweet et la date de soumission de la requête. Ainsi plus la différence est petite, plus le tweet a de chance d'être pertinent vis-à-vis de la requête.

Ainsi ces facteurs ont été combinés linéairement et ont pu montrer des résultats encourageants même s'ils ont montré quelques insuffisances.

3.2 Approches exploitant la structure du réseau social

D'autres travaux ont essayé de répondre à la problématique de la RI dans les microblogs d'un autre angle, ainsi (Nagmoti et al[29]. 2010 ; Das Sarma et al. 2010) ont considéré l'importance des microbloggers comme premier pas pour accéder une information pertinente. De ce fait,le classement des tweets est lié au classement des microbloggers.

Des travaux proposent d'exploiter le réseau social afin d'identifier les acteurs susceptibles de produire du contenu pertinent selon une thématique donnée,

étant donné que le microblogging est une forme de réseau social, il est ainsi possible de traiter le problème de tri des microblogs en exploitant un critère particulier, à savoir le réseau social des plateformes. Cette catégorie d'approches considère que la pertinence est liée à la crédibilité de la source d'information.

Ainsi ces approches œuvrent dans l'identification des microbloggers influents au sein de la plateforme.

Identification des microbloggers influents

Selon Wikipédia, l'influence sociale se produit lorsque les pensées ou les actions d'un individu sont affectées par d'autres individus.

Dans le contexte des microblogs, l'influence est liée à la position de l'utilisateur dans le réseau social.

Le réseau social d'information dans Twitter

Le réseau social de Twitter ne se limite pas aux blogueurs et aux relations d'abonnement, il inclut également tous les acteurs et les données qui interagissent entre eux dans les deux contextes de publication et d'utilisation des articles.

Les microbloggers et les tweets sont connectés par différents types de relations, ces relations sont regroupées en quatre catégories :

— Acteur – Acteur : qui est représenté par la relation d'abonnement entre microbloggers.
— Acteur-Donnée : qui est représenté par la publication de tweet par un microblogger.
— Donnée-Donnée : qui est représenté par le retweet, le tag et le partage.
— Donnée-Acteur : qui est représenté par la mention d'un microblogger dans un tweet.

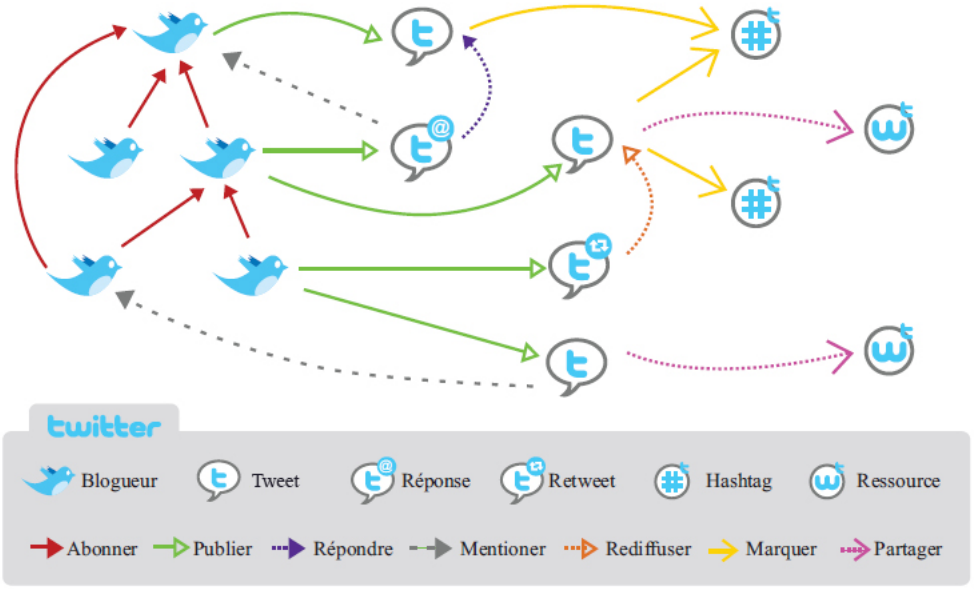


FIGURE 3.1 – Réseau social d'information de Twitter

Modèle de recherche social basé sur l'influence des blogueurs

Un modèle social basé sur l'influence des blogueurs pour la recherche d'information dans les microblogs a été proposé par Benjabeur[17], c'est un modèle de recherche des tweets qui combine la pertinence thématique et l'importance sociale des blogueurs.

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

- **Le réseau social d'influence :** Le réseau social d'influence est extrait à partir du réseau social d'information, ce réseau représente les microbloggers ainsi que les relations sociales entre eux.

Afin d'évaluer l'influence, Benjabeur propose de modéliser le réseau social de Twitter en se basant sur les relations de rediffusion.
En effet, lorsqu'un utilisateur retweet un article, il confirme l'importance du message communiqué. Il montre également qu'il s'intéresse à son sujet et qu'il adopte la même idée si une opinion y est exprimée.
Il est constaté que les abonnés continuent à rediffuser les messages s'ils jugent leur contenu est important.
L'influence d'un blogueur est alors déterminée par la proportion de ses messages rediffusés. Cela exprime aussi son pouvoir d'influence sur le réseau social.
Le réseau social d'influence est modélisé par un graphe $G = (U, E)$ où U est l'ensemble des utilisateurs et $E = U \text{x}\, U$ représente l'ensemble des relations d'influence entre eux. Une relation d'influence $e(u_i, u_j) \in E$ est définie de $u_i \in U$ vers $u_j \in U$ si et seulement s'il existe au moins un article publié par u_j et rediffusé par u_i
le poids $w(u_i, u_j)$ de la relation d'influence est calculé par la formule :

$$w(u_i, u_j) = \frac{\texttt{nb articles publiés par} u_j \texttt{et rediffusés par} u_i}{\texttt{nb articles rediffusés par} u_i} \tag{3.1}$$

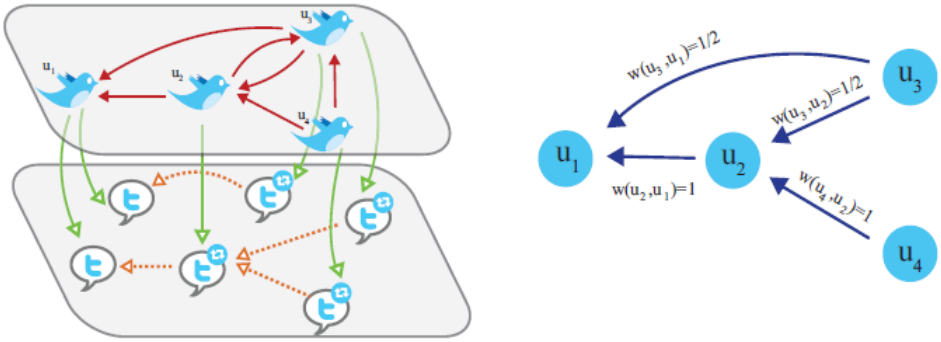


FIGURE 3.2 – Extraction du réseau social d'influence

La recherche des tweets a pour objectif de sélectionner les articles pertinents en réponse à une requête utilisateur.

Benjabeur a proposé un modèle qui combine un score de pertinence thématique et un score de pertinence sociale, afin de présenter une liste précise d'articles.

$$Rel(Q, t, G) = \alpha RSV(Q, , t) + (1 - \alpha) S(Q, u_t, G) \tag{3.2}$$

Où Q, t et G représentent respectivement la requête, l'article et le réseau d'influence social. $\alpha \in [0..1]$ est un paramètre de pondération. $RSV(Q, t)$ est le score normalisé de la

pertinence thématique. $S(Q, ut, G)$ est le score normalisé de l'importance sociale avec ut qui correspond au blogueur ayant publié l'article t.

L'objectif de cette combinaison est de présenter une liste d'articles qui couvrent le sujet de la requête et qui sont publiés par des blogueurs importants.

L'importance sociale d'un blogueur dans ce cas est calculée en fonction de son influence et de son expertise.

$$S(Q, u, G) = InfG(u) * Exp(Q, u) \tag{3.3}$$

Où $InfG(u)$ mesure l'influence du blogueur u dans le réseau social G et le score $Exp(q, u)$ mesure son expertise au sujet de la requête Q.

— **L'influence des bloggeurs**
Le score d'influence est calculé par l'application de l'algorithme PageRank sur le réseau social d'influence. Par analogie avec le principe d'autorité des pages Web, l'influence d'un blogueur provient de l'influence des autres blogueurs avec lesquels il partage une relation d'influence. Plus ces blogueurs sont importants et reçoivent à leur tour des liens d'influence, plus le blogueur est considéré influent. Le score d'influence d'un blogueur est calculé avec la formule suivante :

$$Inf_g(u_i) = d\frac{1}{\mid U \mid} + (1-d) \sum_{u_j : e(u_j,u_i) \in E} w(u_j, u_i) \frac{Inf_g(u_j)}{O(u_j)} \tag{3.4}$$

Avec $O(uj)$ est le nombre de relations d'influence à partir de l'utilisateur $u_j, d \in [0, 1]$ est un paramètre de configuration de l'algorithme PageRank et $w(u_j, u_i)$ est le poids de la relation d'influence entre u_j et u_i comme défini dans l'équation 3.1.

— **L'expertise des bloggeurs**
afin d'évaluer l'expertise d'un blogueur, chaque blogueur va être représenté par un profil B_u, celui-ci correspond à une collection qui contient l'ensemble des microblogs publiés et inclut toutes ses interactions avec la plateforme.
Un modèle de langue est appliqué à cette collection, et un tri des profils est effectué selon leur pertinence à une requête donnée.
Le score obtenu de chaque profil correspond à l'expertise du blogueur par rapport à la requête
Ce score est calculé selon la formule suivante :

$$Exp(q, u) = \prod_{t \in q} ((1-\lambda) p(t \mid B_u) + \lambda p(t))^{n(t,q)} \tag{3.5}$$

$p(t|B_u)$ Correspond à la probabilité d'apparition d'un terme t dans le document

B_u et $p(t)$ est la probabilité d'apparition d'un terme t dans la collection.
$n(t, q)$ représente le nombre d'occurrences du terme t dans la requête q.

Ce modèle a la spécificité d'intégrer la pertinence thématique et la pertinence sociale des tweets. Cette dernière est estimée par la combinaison de deux scores d'influence et d'expertise de chaque blogueur. L'évaluation expérimentale menée montre que cette combinaison permet de mieux évaluer l'importance sociale des blogueurs.

3.3 Approche de reclassement des tweets

Une approche (Masaki Aono [1]) propose de reclasser les résultats obtenus lors d'une recherche effectuée avec un modèle de RI classique en utilisant différents critères.

Cette approche consiste à extraire des critères spécifiques du tweet, puis les combiner au score d'appariement requête-document. Et enfin, classer ces documents selon ce score.

Afin de reclasser les tweets retournés lors de la recherche avec un modèle classique, ils ont proposé un modèle linéaire qui combine les valeurs des différents critères de pertinence de chaque tweet et la pertinence thématique afin d'en estimer la pertinence.

Pour une requête Q et un document T, le score de pertinence $Score(Q, T)$ est estimée par la formule :

$$Score(Q, T) = RSV(Q, T) + \sum_{i=1}^{N} f_i(Q, T) \tag{3.6}$$

Où N est le nombre de facteurs de pertinence et $RSV(Q, T)$ la pertinence thématique.
Ce modèle utilise un groupe de facteurs pertinence dont la plupart se basent sur les statistiques des tweets et des utilisateurs.

— **Les facteurs de pertinence utilisés :**

— Le nombre de retweet : un tweet à fort caractère informatif est retweeté par d'autres utilisateurs.
Retweet count (RTCount) indique le nombre de fois un tweet est retweeté. Ainsi pour mesurer la popularité d'un tweet. Un paramètre RTP dont la valeur est

comprise entre 1 et 5 est introduit :

$$RTP = \begin{cases} 0, & \text{si RTCount} == 0 \\ 1, & \text{si RTCount} \in [1, 10] \\ 2, & \text{si RTCount} \in [11, 100] \\ 3, & \text{si RTCount} \in [101, 1000] \\ 4, & \text{si RTCount} \in [1001, 10000] \\ 5, & \text{pour les autres valeurs} \end{cases}$$

— Le nombre de followers :
FCount indique le nombre de followers que l'auteur d'un tweet a , pour mesurer la crédibilité d'un auteur un paramètre FCP dont la valeur est comprise entre 1 et 5 a été introduit.

$$FCP = \begin{cases} 0, & \text{si FCount} == 0 \\ 1, & \text{si FCount} \in [1, 10] \\ 2, & \text{si FCount} \in [11, 100] \\ 3, & \text{si FCount} \in [101, 1000] \\ 4, & \text{si FCount} \in [1001, 10000] \\ 5, & \text{pour les autres valeurs} \end{cases}$$

— Le nombre de publications d'un auteur :
le paramètre status count d'un tweet , indique le nombre de tweets que l'auteur a publiés avant la publication du dit tweet. Le paramètre SCP est mesuré de la manière suivante :

$$SCP = \begin{cases} 0, & \text{si STCount} == 0 \\ 1, & \text{si STCount} \in [1, 10] \\ 2, & \text{si STCount} \in [11, 100] \\ 3, & \text{si STCount} \in [101, 1000] \\ 4, & \text{si STCount} \in [1001, 10000] \\ 5, & \text{pour les autres valeurs} \end{cases}$$

— le facteur temporel :
C'est un facteur qui mesure la proximité temporelle entre la date de publication du tweet et celle de soumission de la requête, elle est mesurée par la formule :

$$TempScore = \frac{1}{\sqrt{Query_{Time} - Tweet_{time} + 1}}$$

L'exploitation de ces facteurs de façon individuelle lors des expérimentations a

légèrement amélioré les résultats de recherche, mais en les combinant les résultats sont beaucoup plus satisfaisants.

3.4 Approche de classement par modèle d'apprentissage

Duan et Al[9], ont proposé une méthode pour classer les tweets par rapport à leur pertinence par rapport à la requête, ils ont examiné l'effet de certains critères afin de produire un système de classement basé sur l'approche "Learning to Rank".

Ils ont adopté une approche basée sur les algorithmes "Learning to Rank" qui ont déjà fait preuve d'excellentes performances dans les résolutions de différents problèmes liés aux moteurs de recherche.

L'approche Learning To Rank

Learning to rank ou bien Machine learned rank est une approche se basant sur des algortihmes d'apprentissage pour la construction de modèles de classement pour des systèmes de recherche d'information.

Comme première étape, un corpus d'entrainement (d'apprentissage) est préparé duquel sont tirés les facteurs de pertinence.

un algorithme d'apprentissage est utilisé pour entrainer le modèle de classement en exploitant le corpus d'entrainement. à l'issu de cette étape cet algorithme sélectionne et combine les facteurs de pertinence et compare les combinaisons afin d'avoir les meilleurs résultats possibles.

Finalement, le modèle de classement est évalué en utilisant un corpus de test.

La figure suivante illustre le processus.

L'une des tâches les plus importantes de l'approche Learning To Rank est l'extraction de l'ensemble des facteurs de pertinence. Les facteurs extraits sont classés en trois catégories :

- Les facteurs de pertinence basés sur le contenu : ils réfèrent aux facteurs qui décrivent la pertinence entre la requête et le tweet ainsi que tout ce qui touche au contenu du tweet.
- Les facteurs spécifiques à Twitter : ce sont les facteurs qui décrivent les caractéristiques particulières à Twitter telles que le nombre de retweet ainsi les URL partagées dans les tweets.

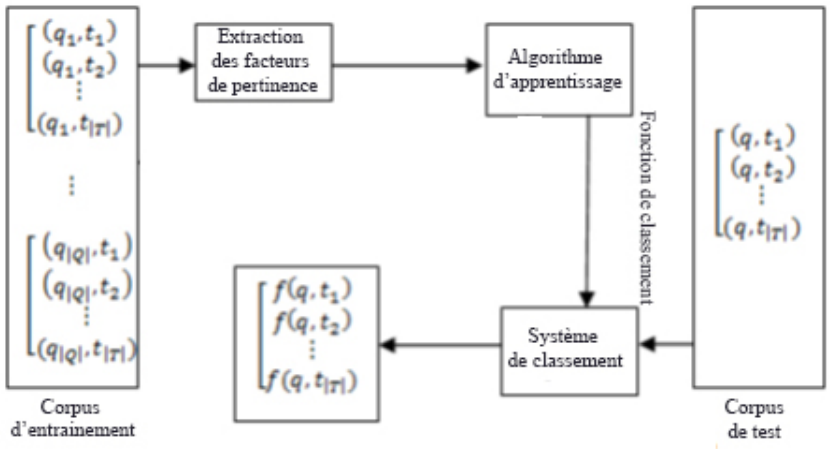


FIGURE 3.3 – l'approche Learning To Rank

— Les facteurs d'autorité : ces facteurs représentent l'influence des auteurs du tweet dans Twitter.

Description des facteurs de pertinence

Facteurs de pertinence basés sur le contenu

Les auteurs de cette approche ont défini trois facteurs de pertinence qui sont : la pertinence thématique mesurée à l'aide du modèle Okapi BM25, la similarité du contenu des tweets et la longueur du tweet.

- Okapi BM25 [1] :
  Ce modèle mesure la correspondance entre la requête et le document.
- Similarité du contenu :
  Ce facteur estime la popularité d'un document dans le corpus (Song et al 2008). Dans le cas de Duan et Al, il mesure le nombre de tweets retournés pour la requête qui sont similaire en contenu au tweet actuel. Pour chaque paire de tweets, le score de similarité se calcule à l'aide du cosinus.
  Le score final de similarité d'un tweet T_i dans T_{Q_k} est calculé selon la formule :

$$Similarite(T_i) = \frac{1}{\mid T_{Q_k} \mid -1} \sum_{T_j \in T_{Q_k}, j \neq i} \frac{TV_i.TV_j}{\mid TV_i \mid . \mid TV_j \mid} \tag{3.7}$$

  Ou TV_i représente le score $TFIDF$ du vecteur T_i et T_{Q_k} représente la collection de tweets retournés pour la requête Q.
- Longueur du tweet :

1. regarder chapitre I : modèle probabiliste

la longueur d'un tweet peut être descriptive de sa valeur en information, ainsi elle peut être utilisée pour mesurer la richesse en informations d'un tweet.

Facteurs spécifiques à Twitter

Les tweets ont beaucoup de caractéristiques, Duan et Al ont décidé d'exploiter certaines de ces caractéristiques :

- URL et URL Count : Un tweet peut contenir une URL , la présence d'une URL est considérée dans cette approche comme un facteur de pertinence, ce facteur est binaire il peut prendre comme valeur 0 ou 1.
  URL Count calcule la fréquence d'apparition d'une URL dans le corpus.
- Fréquence d'un hashtag : pour chaque hashtag apparaissant dans un tweet , un facteur de pertinence qui est égal à la fréquence du hashtag dans le corpus est calculé.
- Tweet Réponse : c'est un facteur qui prend une valeur binaire qui est peut être égale à 1 si le tweet est une réponse à un autre tweet, à 0 sinon.
- OOV (Out of vocabulary) : ce critère est utilisé pour évaluer la qualité de langage dans les tweets, un dictionnaire est utilisé et permet d'identifier les erreurs de langage. Le facteur qualité est calculé comme suit :

$$Qualité(T) = \frac{Nombre(OOV) \in T}{Longueur(T)}$$

Facteurs en rapport avec les auteurs

L'auteur a une place principale au sein du réseau social , de ce fait son importance et la portée de ses publications influent directement sur les résultats de recherche.

L'utilisateur qui a le plus de followers, qui a été mentionné dans plus de tweets et qui a retweeté le plus , est plus important qu'un autre, de ce fait un tweet est plus susceptible de contenir une forte valeur en information s'il a été publié ou retweeté par un utilisateur important.

afin de mesurer l'importance d'un utilisateur, les auteurs ont proposé une variante de l'algorithme PageRank :

$$PScore_{t+1}(v_i) = (1-e) + e. \sum_{v_j \in R_{v_i}} \frac{PScore_t(v_j) RN_{ij}}{N_j} \tag{3.8}$$

avec : R_{v_i} l'ensemble des utilisateurs qui ont retweeté le tweet de v_i.
RN_{ij} est le nombre de fois que v_i a été retweété par v_j N_j le nombre de tweets que v_j a retweeté. et e le facteur d'amortissement.

3.5 Expansion des requêtes et des documents

L'un des plus grands problèmes dans la recherche ad hoc de microblogs est la limite de vocabulaire de chaque document due à leur faible longueur. De ce fait, un ensemble d'approches se basant sur l'expansion des requêtes et des documents ont été proposées afin de pallier à ce problème.

Callan et al[18] (2012) ont proposé une approche qui se base sur deux points principaux :

- l'expansion des documents.
- et l'expansion des requêtes.

L'expansion des documents

Avant d'entamer l'expansion des documents, un pré-traitement est effectué sur les tweets, il comprend :

— La suppression des termes superflus.

— La détection d'URL et le téléchargement de leur contenu.

- **Suppression des termes superflus** :
  Les tweets peuvent contenir des liens ainsi que des caractères spéciaux , même si ces détails peuvent être exploités pour améliorer les résultats de recherche (expansion par source externe), leur indexation avec le contenu textuel du tweet nuit aux résultats de recherche.
  De ce fait, une opération d'élimination de termes superflus est exécutée . à l'issue de cette étape chaque tweet se retrouve dépouillé de son URL, seuls les hashtags sont laissés, les tweets ayant moins de 8 caractères sont jugés non pertinents, car ne contenant pas assez d'informations.
- **Détection d'URL et téléchargement de contenu :**
  Les URLs sont extraites du tweet, et pour chaque URL trouvée dans le corpus , la page web est téléchargée et stockée dans le disque avec une annotation vers l'id du tweet duquel l'URL est extraite, ces pages web sont utilisées ultérieurement pour l'expansion des documents.

Des chercheurs ont remarqué que les tweets pertinents et informationnels contenaient souvent des URLs, De ce fait une méthode a été proposée pour exploiter le document auquel renvoie le lien présent dans le tweet afin de pallier au problème de non-correspondance du vocabulaire.

Généralement, l'expansion de document se fait par la sélection des N termes les mieux

pondérés d'un document HTML, mais cette méthode présente des résultats limités.

De ce fait , les auteurs ont choisi d'utiliser les métadonnées des pages web tels que les termes entre les balises <title> et les mots clés ainsi que les tags. Cette seconde méthode a donné des résultats plus encourageants.

L'expansion des requêtes par réinjection de pertinence

afin de pallier au problème de la mauvaise correspondance de vocabulaire, un modèle utilisant la réinjection de pertinence a été proposé, méthode efficace qui effectue l'expansion de la requête avec des termes extraits des documents pertinents retournés, ces termes sont ensuite ajoutés à la requête originale comme des termes d'expansion.

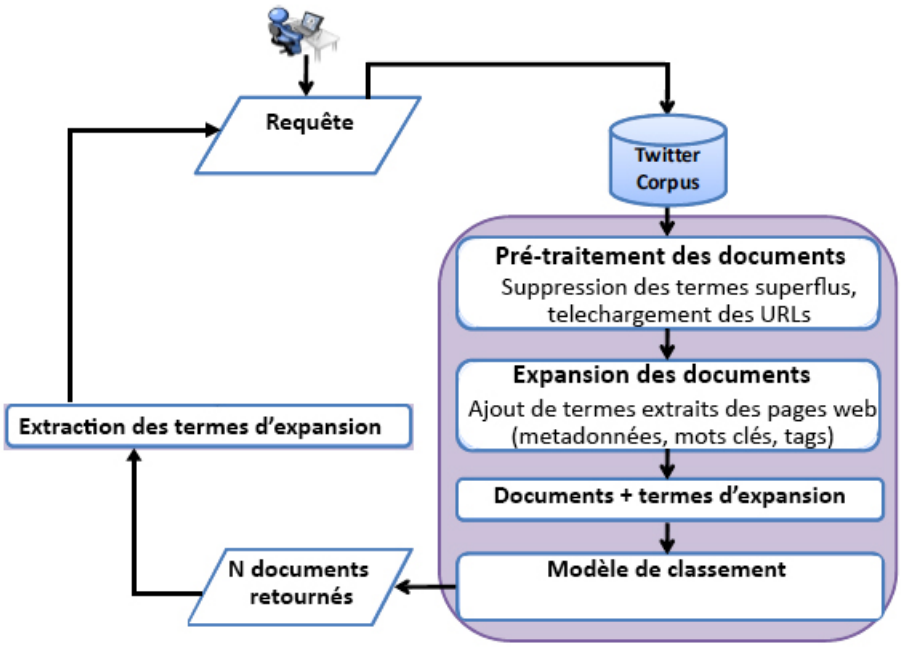


FIGURE 3.4 – Modèle de recherche par expansion de documents et de requêtes

Conclusion

Dans ce chapitre , nous avons fait un état de l'art de la recherche d'information dans les microblogs, nous avons introduit plusieurs approches distinctes qui se basent sur différentes informations disponibles sur la plateforme de microblogging , chacune de ces approches a montré des résultats plus ou moins satisfaisants, dans la partie suivante nous détaillerons l'approche que nous avons proposée.

Deuxième partie

Étude des facteurs de pertinence des microblogs et utilisation des signaux sociaux pour l'amélioration des résultats

Chapitre 4

Integration des signaux sociaux dans le modèle de recherche

Introduction

Dans le chapitre précédent sur l'état de l'art de la recherche d'information dans les microblogs, nous avons passé en revue de nombreuses approches existantes pour la recherche dans les microblogs.

Ces approches proposent d'exploiter plusieurs critères afin de résoudre cette problématique, certaines se basent sur l'étude du réseau social et les interactions entre les individus et d'identifier les acteurs influents et susceptibles de produire des informations pertinentes (Ben jabeur [17]), ou sur l'expansion des documents et des requêtes(Callan [18]) afin d'aboutir à des résultats plus précis, tandis que d'autres approches proposent de classer les documents de sorte à ce que les documents les plus pertinents soient retournés en premier (Masaki Aono[1]).

nos travaux portent sur la proposition d'une approche de RI dans twitter qui permet d'allier efficacité et simplicité de mise en œuvre. Dans ce contexte nous nous basons sur l'approche de Aono[1] que nous proposons d'améliorer.

4.1 Critique de l'approche de Masaki Aono

Les travaux de Masaki Aono[1] (cités précédemment dans le chapitre sur l'état de l'art)ont montré des insuffisances que nous énumérons dans ce qui suit :

Nous regroupons nos reproches dans les points suivants :

- Le facteur temporel est inclus dans les facteurs sociaux, alors que le facteur temporel est indépendant du tweet et est lié au temps de soumission de la requête.
- Une même importance est accordé à tous les critères sociaux pris en compte ce qui

n'est pas forcément représentatif de la pertinence sociale d'un tweet, par exemple, le nombre de followers d'un auteur est plus représentatif de sa popularité que le nombre de statuts qu'il a publiés.

- les scores des facteurs de pertinence sont calculés selon une fonction de distribution discrète, ce qui ne traduit pas fidèlement le score de chaque facteur, par exemple : un tweet ayant 1200 retweets , et un autre qui a 9800 retweets auront le même score selon cette approche.

Dans notre approche nous tentons de remédier à ces insuffisances, en proposant des améliorations que nous allons expliquer dans ce qui suit.

4.1.1 Solutions proposées

Afin d'améliorer l'approche de Aono, nous avons décidé d'apporter les modifications suivante :

- Le facteur temporel sera combiné indépendamment des facteurs sociaux.
- Dans le calcul de la pertinence sociale nous introduirons des coefficients d'amortissement qui permettront d'ajuster le poids de chaque facteur, (ainsi nous pouvons donner plus d'importance à un facteur par rapport à un autre).
- nous avons proposé une fonction de distribution continue pour calculer les facteurs de pertinence sociale.

4.1.2 Approche proposée

Notre approche s'inspire de celle de Masaki Aono qu'elle propose d'améliorer, nous partons donc de l'hypothèse qu'un tweet doit être classé en fonction de (1) sa pertinence thématique,(2) sa pertinence sociale et (3) sa pertinence temporelle.

Le rang d'un tweet dans le classement global des résultats sera donc déterminé par son score global, qui est calculé par la combinaison des valeurs des différents facteurs, notre but est de fournir des résultats aussi pertinent que possible et aussi frais que possible.

Ainsi Notre fonction de classement est définie comme suit :

$$ScoreClassement(Q, T_i) = \alpha RSV(Q, T_i) + \beta ScoreTwitt(T_i) + \gamma Temp(Q, T_i) \quad (4.1)$$

Avec $RSV(Q, T_i)$ la pertinence thématique, $ScoreTwitt(T_i)$ le score social du tweet et $Temp(Q, T_i)$ la pertinence temporelle(ces facteurs seront expliqués en détails ultérieurement).

α, β,γ des coefficients d'amortissement calculés de manière expérimentale, ils permettent d'équilibrer entre les différents facteurs en ajustant leur poids dans l'équation.

Calcul des facteurs de pertinence

- **Facteur temporel**
  Nous prenons en compte à ce niveau le facteur fraîcheur par rapport à la date de soumission de la requête dans la mesure de la pertinence. Ce facteur permet d'amplifier les scores de pertinence du contenu d'un tweet en fonction de sa proximité temporelle avec la date de la requête.
  le tweet le plus récent aura le plus haut score de fraicheur, ce qui augmente la probabilité que les tweets les plus hauts classés soient récents.

$$TempScore(Q, T_i) = \frac{1}{\sqrt{Q_{Time} - T_i time + 1}} \tag{4.2}$$

- **Facteur de pertinence thématique (similarité requête-document)**
  ceci refere au score de correspondance des termes de la requête avec les documents. Il peut être calculé de différentes manières selon le modèle utilisé. dans notre cas nous avons utilisé le modèle vectoriel VSM.
  Les tweets ayant un score de similarité haut sont plus pertinents du point de vue thématique.

- **Facteurs en rapport avec le tweet :**
  Les facteurs liés au tweet incluent :
  - L'impact du tweet : le tweet a-t-il été retweeté, si oui, combien de fois ?
  - Le contenu : le tweet contient-il des liens hypertextes ?
  - Le degré d'activité de l'auteur : l'auteur est-il actif, combien de tweets a-t-il publiés ?
  - La popularité de l'auteur : l'auteur est-il populaire au sein du réseau ?

  Calcul des facteurs liés au tweet :

  — L'impact du tweet :
  Dans cette étape, nous prenons en compte le nombre de fois qu'un tweet en particulier a été retweeté par les utilisateurs au sein de Twitter, ainsi le tweet qui est le plus retweeté aura le plus haut score étant donné qu'il est considéré comme étant le plus populaire.

$$f_1(T_i) = Impact(T_i) = log_{10}(RT_{T_i}) \tag{4.3}$$

  — - Présence d'URLs dans le tweet :

Dans cette étape, on vérifie si le tweet contient un lien hypertexte ou n'importe quel autre média, les tweets ayant ce type de contenu auront un plus grand score car ils sont susceptibles d'avoir une plus grande valeur informative.

$$f_2(T_i) = \begin{cases} 1 & \text{si } T_i \text{ contient une URL} \\ 0 & \text{sinon} \end{cases} \tag{4.4}$$

— La popularité de l'auteur du tweet :
Le nombre de followers d'un utilisateur peut être considéré comme un indice de popularité de celui-ci.

$$f_3(T_i) = Popoluarite(AuteurT_i) = log_{10}(NbrFollowers) \tag{4.5}$$

— Le degré d'activité de l'auteur :
Ce facteur prend en compte le nombre de publications de l'auteur, son objectif est de valoriser les tweets publiés par des auteurs actifs par rapport aux tweets publiés par des auteurs moins actifs.

$$f_4(T_i) = Active(AuteurT_i) = log_{10}(Nbrstatuts) \tag{4.6}$$

- Score du tweet :

Tous les facteurs en rapport avec le tweet seront combinés pour calculer le score $ScoreTwitt(T_i)$,la score qui définit l'importance sociale d'un tweet.
Mais avant de combiner le score de ces facteurs, nous aurons besoin de les normaliser, la normalisation est le processus qui permet de réajuster les valeurs des facteurs pour qu'elles se situent dans l'intervalle [0.1], pour y arriver, chaque valeur d'un facteur va être divisée par la valeur maximale qu'il peut atteindre.

$$Norm(f_i) = f_i/Max(f_i)$$

le score $ScoreTwitt(T_i)$ est calculé selon la formule suivante :

$$ScoreTwitt(T_i) = \sum_{i=0}^{i=N} \alpha_i f_i$$

Avec α_i des coefficients d'amortissement qui permettent d'ajuster le poids de chaque facteur, ces coefficients seront calculés expérimentalement.

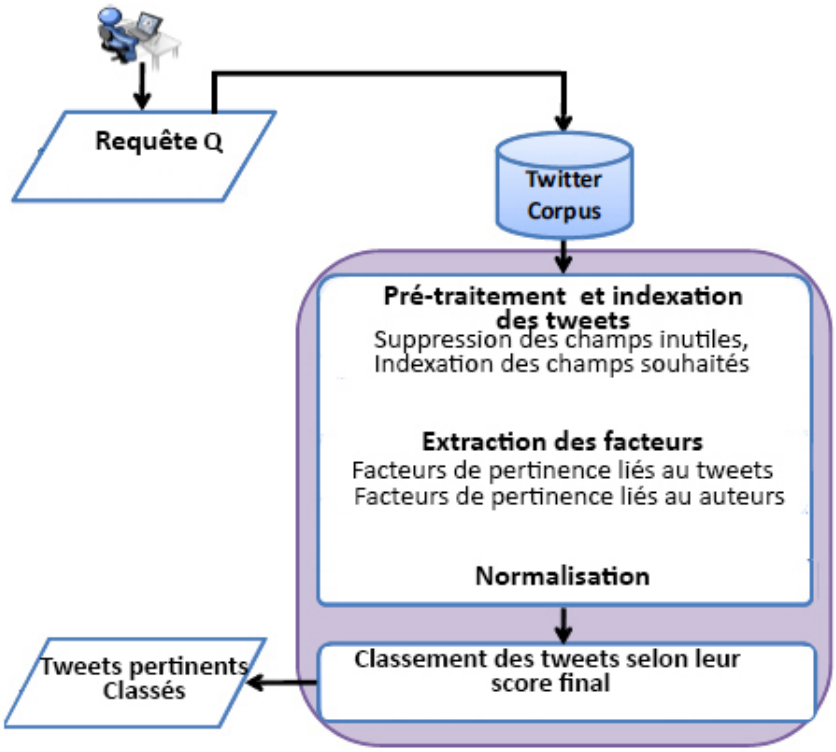


FIGURE 4.1 – Processus de recherche des tweet

Conclusion

Dans ce chapitre nous avons décrit en détail notre approche ainsi que tout ce qu'elle a apporté comme améliorations à celle de Masaki Aono nous avons exploré en détail chaque facteur de pertinence utilisé par notre approche, nous avons décrit toutes les formules proposées dans notre méthode, dans le prochain chapitre, nous allons voir l'implémentation de notre approche ainsi que l'évaluation de ses résultats.

Chapitre 5

Implémentation et expérimentations

Introduction

Dans ce chapitre, nous présentons les expérimentations effectuées pour évaluer l'apport des différentes propositions faites dans le chapitre précédent.

Dans ce qui suit nous présentons le cadre expérimental et nous décrivons la collection de données ainsi que les mesures d'évaluation utilisées.

Nous allons ensuite comparer les résultats de notre approche avec les résultats de celle dont elle est inspirée ,ceci dans le but de déterminer la fiabilité et la performance de notre approche.

Enfin, nous allons faire part des conclusions auxquelles nous avons abouties par l'étude et l'analyse des résultats de notre approche.

5.1 Implémentation

Nous avons implémenté notre approche sous Java en utilisant la bibliothèques Lucene, nous avons développé ce projet sous l'IDE eclipse.

Dans notre implémentation nous avons extrait les facteurs de pertinence à travers l'API twitter [1].

5.2 Cadre expérimental

5.2.1 Protocole expérimental

Notre évaluation va se faire comme suit : nous allons effectuer une série de test avec notre approche puis nous comparerons ceux-ci avec les résultats retournés par l'approche

1. voir l'annexe

de Masaki Aono que nous avons implémentée.

Pour rappel, Notre approche se base sur sur le modèle vectoriel comme modèle de restitution , notre fonction de classement est sous la forme :

$$ScoreClassement(Q, T_i) = \alpha RSV(Q, T_i) + \beta ScoreTwitt(T_i) + \gamma Temp(Q, T_i) \quad (5.1)$$

Avec $\alpha = 0.5$, $\beta = 0.3$ et $\gamma = 0.2$ ces valeurs sont les plus optimales que nous avons déterminé après des séries de tests.

L'approche de Masaki Aono se base sur le même modèle de restitution, la fonction de classement est définie comme suit :

$$Score(Q, T) = RSV(Q, T) + \sum_{i=1}^{N} f_i(Q, T) \quad (5.2)$$

5.2.2 La collection de test

Toutes les expérimentations réalisées dans ce chapitre ont été effectué sur la collection définie dans ce qui suit : La collection de test utilisée est constituée de tweets extraits à l'aide de l'API twitter , cette collection inclut 502 tweets qui se situent entre Mai 2010 et Mars 2011, un ensemble 802 de requêtes et un ensemble de 802 jugement de pertinence.

5.2.3 Mesures d'évaluation utilisées

Pour estimer la qualité des listes de résultats produites selon les différentes approches, nous avons utilisé les mesures standards de la précision moyenne (Average précision), la MAP, la R-Précision et Précision@X et la F-mesure.

5.2.4 Résultats

Comparaison entre nos résultats et ceux de Masaki Aono

Dans cette section, nous présentons les résultats obtenus lors des expérimentations de notre approche et celle de Masaki Aono (considérée comme Baseline) que nous avons implémentée sous Lucene, puis nous comparons ces résultats et nous les analyserons. Nous récapitulons les résultats dans les tableaux et graphiques suivants :

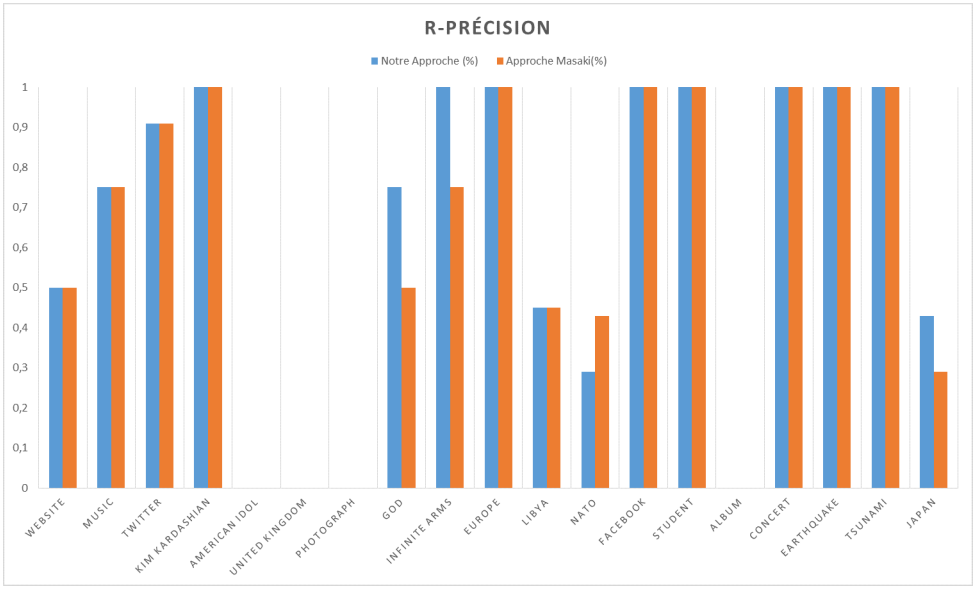


FIGURE 5.1 – Comparaison de la précision réele entre les deux approches

		R-Précision	
Identifiant de la Requête	intitulé de la Requête	Notre Approche	Approche Masaki
33898	Website	0,5	0,5
18839	Music	0,75	0,75
9988187	Twitter	0,909090909	0,909090909
19394613	Kim Kardashian	1	1
191890	American Idol	0	0
31717	United Kingdom	0	0
25080	Photograph	0	0
5042765	God	**0,75**	**0,5**
26475979	Infinite arms	**1**	**0,75**
9239	Europe	1	1
17633	Libya	0,45	0,45
21133	NATO	0,29	0,43
7529378	Facebook	1	1
155526	Student	1	1
528282	Album	0	0
236918	Concert	1	1
10106	Earthquake	1	1
31161	Tsunami	1	1
15573	Japan	**0,43**	**0,29**

TABLE 5.1 – La précision réelle des deux approches

Nous remarquons que sur la majorité des requêtes, la précision réelle de notre approche est égale ou supérieure à la baseline notamment pour les requêtes 26475979 , 5042765 et 15573 qui ont donné des résultats nettement meilleurs.

		Précision Moyenne	
Identifiant de la Requête	intitulé de la Requête	Notre Approche	Approche Masaki
33898	Website	**0,4166**	**0,375**
18839	Music	0,6042	0,6042
9988187	Twitter	0,6634	0,6634
19394613	Kim Kardashian	0,6667	0,6667
191890	American Idol	0	0
31717	United Kingdom	0	0
25080	Photograph	0	0
5042765	God	**0,4792**	**0,3583**
26475979	Infinite arms	1	0,6792
9239	Europe	0,6667	0,6667
17633	Libya	**0,7075**	**0,6548**
21133	NATO	0,2996	0,4176
7529378	Facebook	1	1
155526	Student	0,6667	0,6667
528282	Album	0,2167	0,2167
236918	Concert	0,25	0,25
10106	Earthquake	0,6	0,6
31161	Tsunami	0,6667	0,6667
15573	Japan	**0,5818**	0,3429

TABLE 5.2 – La précision Moyenne des deux approches

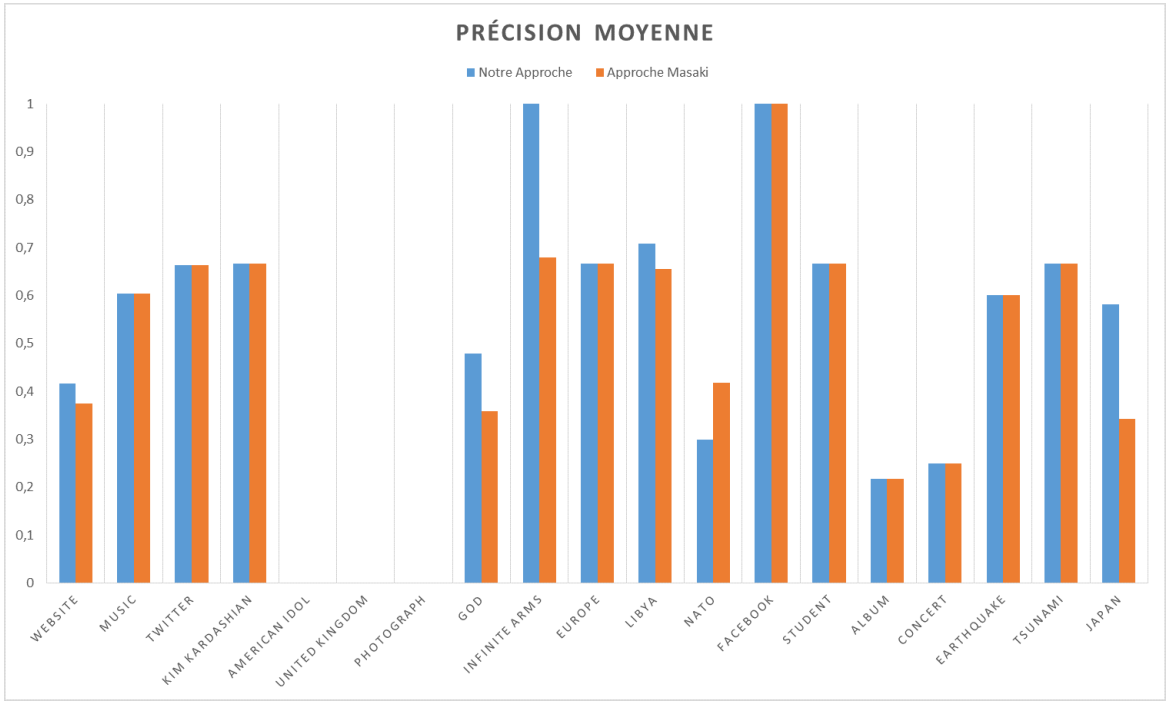


FIGURE 5.2 – Comparaison de la précision moyenne entre les deux approches

Sur la majorité des requête notre approche à donné des résultats similaires ou meilleurs par rapport à la baseline, notamment pour les requêtes 33898 , 5042765, 26475979 , 17633 , 15573, celles-ci ont montré des résultats supérieurs à la baseline.

		Précision à 5 documents	
Identifiant de la Requête	intitulé de la Requête	Notre Approche	Approche Masaki
33898	Website	**0,4**	**0,2**
18839	Music	0,6	0,6
9988187	Twitter	1	1
19394613	Kim Kardashian	1	1
191890	American Idol	0	0
31717	United Kingdom	0	0
25080	Photograph	0	0
5042765	God	**0,6**	**0,4**
26475979	Infinite arms	0,8	0,8
9239	Europe	1	1
17633	Libya	**0,8**	**0,6**
21133	NATO	0,2	0,6
7529378	Facebook	1	1
155526	Student	0,4	0,4
528282	Album	0,4	0,4
236918	Concert	1	1
10106	Earthquake	1	1
31161	Tsunami	1	1
15573	Japan	**0,6**	**0,2**

TABLE 5.3 – La précision à 5 documents des deux approches

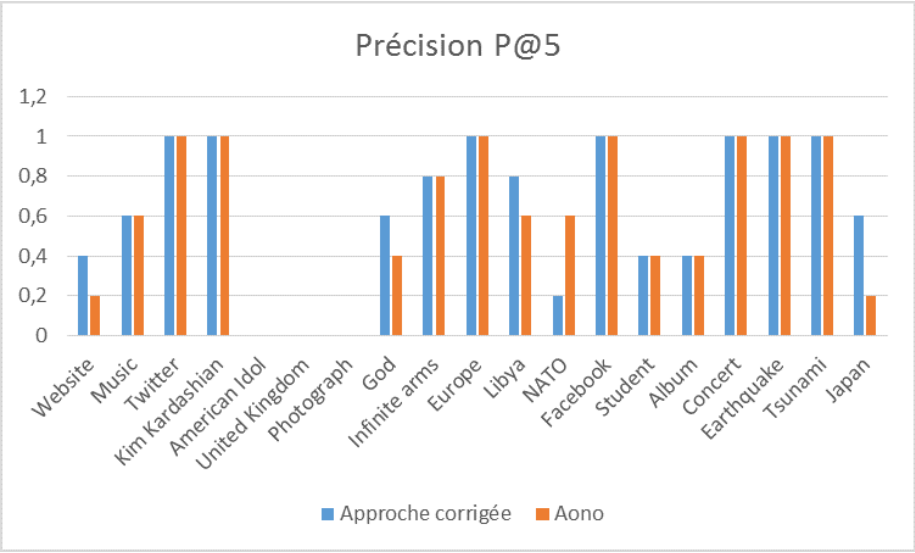


FIGURE 5.3 – Comparaison de la précision à 5 documents entre les deux approches

D'après les résultats retournés, notre approche a une meilleures précision aux cinq premiers documents retournés , notamment pour les requêtes 5042765, 17633,33898 et 15573 qui ont retourné des résultats nettement supérieurs à la baseline.

	(Résultats)	
Mesure	Notre Approche	Approche Masaki
F-mesure	0,31167929	0,31167929
MAP	0,499252632	0,465778947
P@5	0,621052632	0,589473684

TABLE 5.4 – Comparaison des performances des deux approches

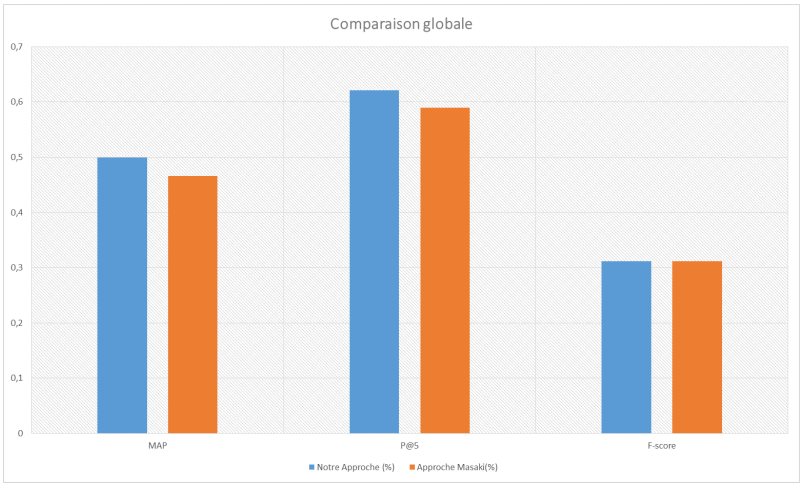


FIGURE 5.4 – Comparaison des résultats globaux entre les deux approches

Notre approche améliore nettement les résultats selon les métriques utilisées , ainsi nous avons obtenu une MAP et une P@5 supérieure à la baseline, nous avons obtenu un taux d'amélioration de 3.4% pour la MAP. La F-mesure est la même pour les deux approches, nous concluons donc que notre approche n'augmente pas le nombre de documents pertinents restitués , mais augmente la vitesse de restitution , en effet notre approche améliore le classement des documents pertinents.

Synthèse

Après de nombreuses séries de test et d'expérimentations puis analyse des résultats , nous avons conclu que notre approche apportait de nettes améliorations à celle de Masaki Aono , la collection sur laquelle nous avons effectué nos expérimentations n'était pas volumineuse mais malgré celà ,nous avons obtenu des résultats satisfaisants ,cela démontre la pertinence de nos propositions et nous encouragent dans notre démarche.

Conclusion

Dans ce chapitre nous avons présenté le cadre expérimental de nos travaux, puis nous avons réalisé des expérimentations sur un corpus de test, nous avons analysé ces résultats et nous les avons comparés aux résultats de l'approche de Masaki Aono. Nous sommes arrivés à la conclusion que notre approche apporte certaines améliorations dans ce cadre.

Nous avons trouvé que notre approche améliore les résultats de certaines requêtes ainsi que les performances de recherche globale , Ceci démontre la pertinence de notre approche et nous encourage à améliorer notre approche afin d'aboutir à de meilleurs résultats.

Conclusion

Conclusion

Les microblogs ont été un sujet de recherche ces dernières années, ceci dû à la popularité des plateformes de microblogging. Twitter,étant la plateforme qui a connu le plus de croissance ces dernières année, a suscité l'intérêt de plusieurs chercheurs. Sur Twitter, un utilisateur partage ses idées, ses opinions ainsi que des nouvelles sous forme de courts messages. La popularité grandissante des services de microlobbing, a incité les utilisateurs à l'utiliser comme moyen de satisfaire leurs besoins en information. Ce qui est a poussé les chercheurs à s'intéresser aux problématiques de la recherche d'information dans les microblogs.
Les travaux présentés dans ce manuscrit s'inscrivent dans le contexte de la recherche d'informations dans les microblogs , qui correspond à un des domaines émergents de la RI avec de nombreux enjeux.

Nous avons axé notre revue de l'état de l'art selon cette dimension en donnant un aperçu des différentes approches qui ont été proposées.

Notre objectif était de proposer une approche qui améliore la recherche d'information dans les microblogs.

Pour y arriver, nous avons exploiter les propriétés présentes dans les tweets,, notre approche propose d'exploiter la pertinence thématique du tweet ,ses propriétés "sociales" ainsi que sa fraicheur afin d'atteindre ce but.

Nous avons proposé un modèle pour la recherche sociale des tweets. Ce modèle a la spécificité d'intégrer la pertinence thématique , la pertinence sociale et la pertinence temporelle des tweets.L'évaluation expérimentale préliminaire que nous avons menée sur la collection permet d'aboutir à des résultats. Néanmoins les améliorations infimes que nous avons notés ne permettent pas d'évaluer totalement les performances de cette approche.

Limites et Perspectives

Notre étude a montré quelques limites :

— Nous n'avons pas pu tester chaque facteur de pertinence individuellement.

— La collection sur laquelle nous avons testé notre approche ne contenait pas assez de documents, ce qui n'a pas donné des résultats très claires à interpréter.

Perspectives

En perspective, nous envisageons de mener nos expérimentations sur une collection d'articles et de requêtes de plus grande taille.
Nous voulons tenter d'autres approches telles que l'analyse des réseaux sociaux ou l'expansion de requêtes et de documents, ou bien nous allons étudier l'apport de nouveaux facteurs de pertinence dans l'amélioration des résultats.

Bibliographie

[1] Masaki Aono, Abu Nowshed Chy, and Md Zia Ullah. A time and context aware re-ranker for microblog retrieval. *The 29th Annual Conference of the Japanese Society for Artificial Intelligence*, 2015.

[2] Ismail Badache and Mohand Boughanem. Exploitation de signaux sociaux pour estimer la pertinence a priori d'une ressource. *Conférence francophone en Recherche d'Information et Applications*, 2013.

[3] Mohamed Reda BouadjeneK. *Infrastructure and Algorithms for Information Retrieval Based On Social Network Analysis/Mining.* PhD thesis, 2013.

[4] Meeyoung Cha, Hamed Haddadi, Fabricio Benevenuto, and Krishna P. Gummadi. Measuring user influence in twitter : The million follower fallacy. *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media*, 2011.

[5] Ed H. Chi and Rowan Nairn. Information seeking with social signals : Anatomy of a social tag-based exploratory search browser. *IUI Conference workshop on Social Recommender Systems*, 2010.

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

[7] Houssem Eddine Dridi. Analyse des données de microblogs. 2012.

[8] Shaoyi Duan. *Personalized Microblog Search on Twitter.* PhD thesis, 2012.

[9] Yajuan Duan, Long Jiang, Tao Qin, Ming Zhou, and Heung-Yeung Shum. An empirical study on learning to rank of tweets. *In Proceedings of the 23rd International Conference on Computational Linguistics*, 2012.

[10] Miles Efron. Information search and retrieval in microblogs. *Journal of the American Society for Information Science and Technology*, 2011.

[11] Tarek Elganainy. *Hyperlink-Extended Pseudo Relevance Feedback for Improved Microblog Retrieval.* PhD thesis, 2014.

[12] Marcus Greenborg. *Finding relevant search results in social networks.* PhD thesis, 2011.

[13] Adrien Guille. *Diffusion de l'information dans les médias sociaux :Modélisation et analyse.* PhD thesis, 2014.

[14] Zhongyuan H., Xuwei L., and Muyun Y. Hit at trec 2012 microblog track. *The Twenty-First Text REtrieval Conference (TREC 2012) Proceedings*, 2012.

[15] Hongzhao Huang. *Modeling heterogeneous networks for information ranking, enrichment and resolution on microblogs.* PhD thesis, 2015.

[16] Lamjed BEN JABEUR. *Leveraging social relevance : Using social networks to enhance literature access and microblog search*. PhD thesis, 2013.

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

[18] Y. Kim, R. Yeniterzi, and J. Callan. Overcoming vocabulary limitations in twitter microblogs. *The Twenty-First Text REtrieval Conference (TREC 2012) Proceedings*, 2013.

[19] Sebastian Marius Kirsch. *Social Information Retrieval*. PhD thesis, 2008.

[20] James Lanagan. *Measuring Author Influence on Information Retrieval*. PhD thesis, 2009.

[21] Cher Han Lau. *Detecting News Topics from Microblogs using Sequential Pattern Mining*. PhD thesis, 2014.

[22] Cher Han Lau, YueFeng Li, and Dian Tjondronegoro. Microblog retrieval using topical features and query expansion. *TREC 2011*, 2011.

[23] Yun LI, Yi GUAN, Xishuang DONG, and Xinbo LV. Language modeling for microblog retrieval : Combine multiple-bernoulli model and temporal prior for tweets rank. *Journal of Computational Information Systems*, 2012.

[24] Bei Li1 and Yanjie Liu. A novel approach for microblog message ranking based on trust model and content similarity. *International Journal of Database Theory and Application*, 2011.

[25] Jimmy Lin, Miles Efron, Yulu Wang, and Garrick Sherman. Overview of the trec 2014 microblog track. *The Twenty-Third Text REtrieval Conference (TREC 2014) Proceedings*, 2014.

[26] Magnani M., Montesi D., and Rossi L. Conversation retrieval for microblogging sites. *Journal Information Retrieval*, 2012.

[27] Metzler.D and Cai .C. Microblog track notebook version. 2011.

[28] Taiki Miyanishi, Kazuhiro Seki, and Kuniaki Uehara. Combining recency and topic-dependent temporal variation for microblog search. *Lecture Notes in Computer Science*, 2012.

[29] Rinkesh Nagmoti and Ankur Teredesai. Ranking approaches for microblog search. *Web Intelligence and Intelligent Agent Technology (WI-IAT), 2010 IEEE/WIC/ACM International Conference*, 2010.

[30] ZAHRA AMIN NAYERI. *A Time-Aware Approach to Improving Ad-hoc Information Retrieval from Microblogs*. PhD thesis, 2014.

[31] Stephen E. Robertson and Steve Walker. Okapi at trec-3. proceedings of the third text retrieval conference. 1994.

[32] Markus Schaal, John O'Donovan, and Barry Smyth. An analysis of topical proximity in the twitter social graph. *4th International Conference, SocInfo 2012, Lausanne, Switzerland, December 5-7, 2012. Proceedings*, 2012.

[33] Cunhui Shi, Bo Xu, Hongfei Lin, and Qing Guo. A time-sensitive model for microblog retrieval. 2012.

[34] Kirchhoff. L.and Stanoevska-Slabeva, Nicolai K., and Fleck. Using social network analysis to enhance information retrieval systems. *Applications of Social Network Analysis*, 2008.

[35] Wartik Steven. *Boolean operations : Information Retrieval Data Structures and Algorithms*. 1992.

[36] Ibrahim Uysal and W. Bruce Croft. User oriented tweet ranking : A filtering approach to microblogs. *Proceedings of the 20th ACM international conference on Information and knowledge management*, 2012.

[37] C. J. van Rijsbergen. *Information Retrieval book*. 1979.

[38] Stanley Wasserman and Katherine Faust. *Social Network Analysis Methods and Applications*. 1995.

[39] Jianshu WENG, Ee Peng LIM, Jing JIANG, and Qi HE. Twitterrank : Finding topic-sensitive influential twitterers. *Proceedings of the Third ACM International Conference on Web Search and Data Mining*, 2011.

[40] Craig Willis, Richard Medlin, and Jaime Arguello. Incorporating temporal information in microblog retrieval. 2012.

Annexe

Outils de développement

Java

Le langage Java est un langage de programmation informatique orienté objet créé par James Gosling et Patrick Naughton, employés de Sun Microsystems, La particularité et l'objectif central de Java est que les logiciels écrits dans ce langage doivent être très facilement portables sur plusieurs systèmes d'exploitation tels que UNIX, Windows, Mac OS ou GNU/Linux, avec peu ou pas de modifications. Pour cela, divers plateformes et frameworks associés visent à guider, sinon garantir, cette portabilité des applications développées en Java.

Lucene

Lucene est un moteur de recherche textuelle Open Source et passant bien à l'échelle fourni par la fondation Apache. Vous pouvez utiliser Lucene dans des applications commerciales ou Open Source. Les puissantes APIs de Lucene se concentrent surtout sur l'indexation et le recherche. Il peut être utilisé pour ajouter des capacités d'indexation à des applications comme des clients de courrier, des listes de diffusion, des applications effectuant des recherches sur Internet ou dans une base de données, etc. Des sites Internet tels que Wikipedia, TheServerSide, jGuru, et LinkedIn utilisent Lucene.

Classes principales de Lucene

Les sections suivantes fournissent une brève introduction aux principales classes qui servent à construire ce moteur de recherche.

- Classes d'indexation

  — IndexWriter - La classe IndexWriter est le composant central du processus d'indexation. Cette classe crée un nouvel index et ajoute 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.

— Directory - La classe Directory représente l'emplacement de l'index de Lucene. IndexWriter utilise une des implémentations de Directory, FSDirectory, pour créer son index dans un répertoire dans le Système de fichiers. Une autre implémentation, RAMDirectory, prend toutes ses données en mémoire. Cela peut être utile pour de plus petits indices qui peuvent être pleinement chargés en mémoire et peuvent être détruits sur la fin d'une application.

— 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.

— 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.

— Field - Chaque document est un index contenant un ou plusieurs champs, inséré dans une classe intitulé Field. Chaque champ (field) correspond à une portion de donnée qui est interrogé ou récupéré depuis l'index durant la recherche.

- Classes de recherche

— 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.

— Term - Un terme est une unité basique pour la recherche, similaire à l'objet field. Il est une chaîne de caractère : le nom du champ et sa valeur. Notez que les termes employés sont aussi inclus dans le processus d'indexation.

— Query - La classe Query est une classe abstraite qui comprend BooleanQuery, PhraseQuery, PrefixQuery, PhrasePrefixQuery, RangeQuery, FilteredQuery, et SpanQuery. 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.

— QueryParser - La classe QueryParser est utilisée pour générer un décompositeur analytique qui peut chercher à travers un index.

— 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.

Eclipse IDE

Eclipse est un environnement de développement (IDE) historiquement destiné au langage Java, même si grâce à un système de plugins il peut également être utilisé avec d'autres langages de programmation, dont le C/C++ et le PHP.

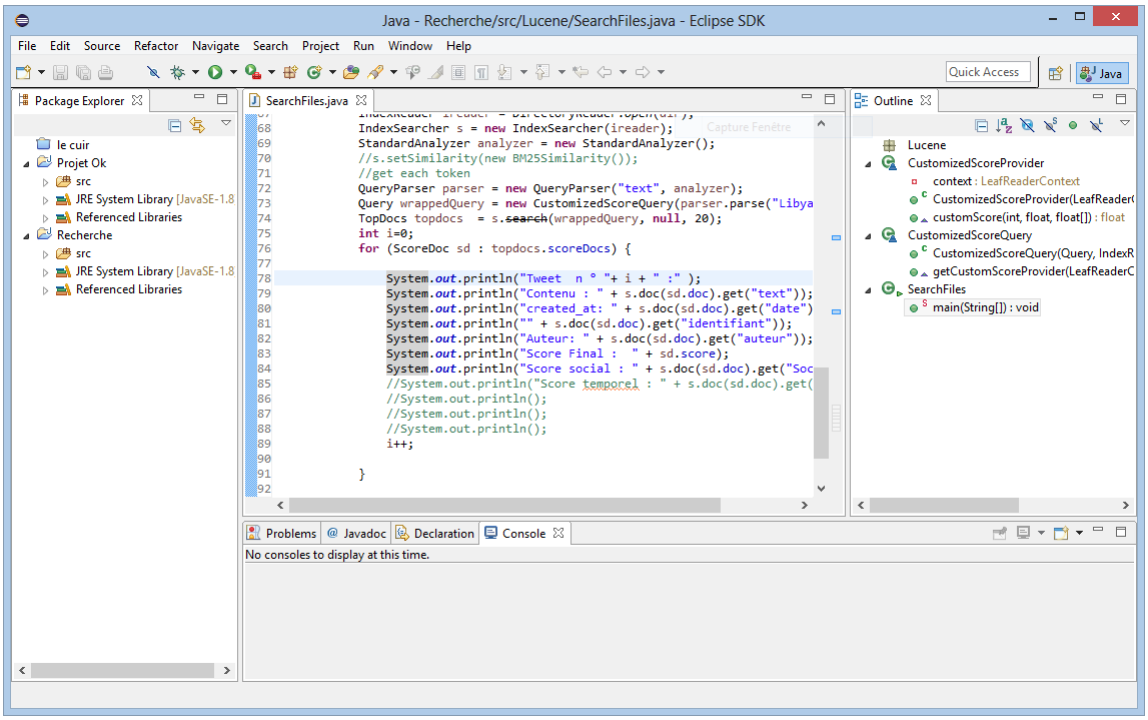


FIGURE 5.5 – Interface d'Eclipse IDE

API Twitter

L'API twitter fournit un ensemble d'informations sur les tweets et les utilisateurs, les définitions suivantes vont nous aider à mieux comprendre certains concepts.

Définitions de termes relatifs à l'API Twitter

Dans ce qui suit, nous allons voir quelques définitions de termes liés à l'API Twitter.

- **Utilisateur(User) :**C'est l'entité qui publie des tweets, suit (follow) d'autres utilisateurs, ils peuvent être mentionnés ou eux même suivis par d'autres utilisateurs, ils ont un ensemble de propriétés qui leur est associés.
  Parmi ces propriétés on peut citer :
  — ID : c'est la représentation numérique de l'identifiant unique d'un utilisateur, il est stocké sur 53 bits.
  — Screen_name : c'est l'alias de l'utilisateur, il est composé d'un maximum de 15 caractères, il est unique pour chaque utilisateur , mais peut être sujet à modification.
  — Friends_count : réfère au nombre d'utilisateurs que le compte actuel suit (ses abonnements).
  — Followers_count : représente le nombre de followers que l'utilisateur a (ses abonnés).

FIGURE 5.6 – Exemple de données de l'API pour un utilisateur

Un tweet a un ensemble de propriétés qui lui sont associées, parmi celles qu'on a utilisées :
— Tweet_id : est un entier qui représente l'identifiant unique d'un tweet, similaire à User_id , celui-ci est stocké sur 64 bits.
— Text : c'est le contenu textuel du tweet (encodé UTF-8).

— Created_on : réfère à la date de création du tweet.
— Retweet_count : représente le nombre de fois qu'un tweet a été partagé par d'autres utilisateurs.

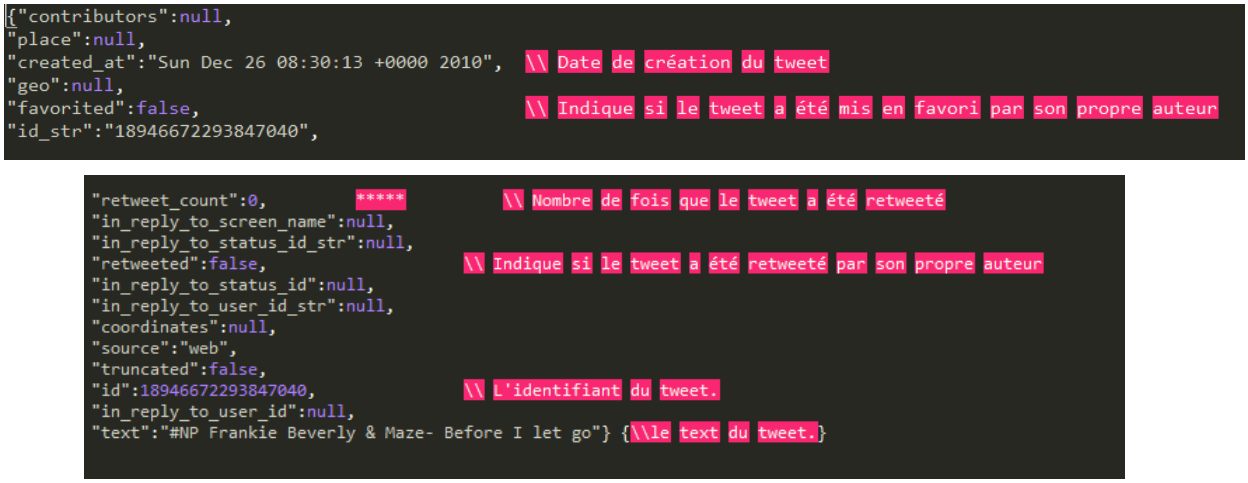


FIGURE 5.7 – Exemples de données de l'API pour un tweet