République **A**lgérienne **D**émocratique et **P**opulaire
Ministère de l'**E**nseignement **S**upérieur et de la **R**echerche **S**cientifique

UNIVERSITE **M**OULOUD **M**AMMERI DE **T**IZI-**O**UZOU

FACULTE DU **G**ENIE **E**LECTRIQUE ET D'INFORMATIQUE
DEPARTEMENT D'INFORMATIQUE

Mémoire de Fin d'Etudes
de MASTER ACADEMIQUE

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

Présenté par
Sabrina BELLOUTI
Malika AMROUNI

Thème

Recherche d'Information Sémantique dans les Microblogs -Cas de Twitter-

Mémoire soutenu publiquement le 22/09/ 2016. Devant le jury composé de :

M. Mohammed Nabil AMIROUCHE	**Président**
Mme Fatiha AMIROUCHE	**Encadreur**
Melle Wassila AZZOUG	**Examinatrice**
Mme Lila BELKACEMI	**Examinatrice**

Remerciements

Tout d'abord, on tient à remercier particulièrement notre encadreur ***Mme AMIROUCHE FATIHA****, de nous avoir fait confiance et encouragé tout au long de ce projet. On la remercie également pour sa disponibilité, son aide, ses conseilles précieux, ses critiques constructives, ses explication et ses suggestions pertinentes.*

On tient aussi à exprimer toutes notre gratitude aux membres du ***jury*** *pour avoir accepté d'évaluer et de juger notre travail.*

J'adresse également mes remerciements à tous les ***professeurs*** *qui nous ont enseigné durant ces cinq dernières années pour le savoir et la formation qu'ils nous ont transmis.*

En fin, nos souhaits les plus chers seront de remercier vivement toutes les personnes, qui, de prés ou de loin, se sont impliquées dans la réalisation de ce projet.

Dédicaces

A cœur vaillant rien d'impossible

A conscience tranquille tout est accessible

Quand il y a la soif d'apprendre

Tout vient à point à qui sait attendre

Tout devient facile pour arriver à nos fins

Malgré les obstacles qui s'opposent

En dépit des difficultés qui s'interposent

Les études sont avant tout

Notre unique et seul atout

Ils représentent la lumière de notre existence

L'étoile brillante de notre réjouissance

Espérant des lendemains épiques

Un avenir glorieux et magique

Souhaitant que le fruit de nos efforts fournis

Jour et nuit, nous mènera vers le bonheur fleuri

Aujourd'hui, ici rassemblés auprès des jurys,

Nous prions dieu que cette soutenance

Fera signe de persévérance

Et que nous serions enchantés par notre travail honoré

Je dédie ce modeste travail à :

A celle qui représente pour moi le symbole de la bonté par excellence, la source de tendresse et l'exemple du dévouement qui n'a pas cessé de m'encourager et de prier pour moi : ma chère ***grande- mère****, Puisse Dieu, le tout puissant, te préserver et t'accorder santé, longue vie et bonheur.*

A ma chère ***mère,*** *A mon cher* ***père,*** *ceux qui ont bercé mes rêves, Aucune dédicace ne saurait exprimer l'amour, l'estime, le dévouement et le respect que j'ai toujours eu pour vous.*

A mon très cher ***mari****, quand je t'ai connu, j'ai trouvé l'homme de ma vie, mon âme sœur et la lumière de mon chemin.*

A mon cher ***frère*** *et ma chère* ***sœur****, les mots ne suffisent guère pour exprimer l'attachement, l'amour et l'affection que je porte pour vous.*

A mes très chères ***tantes et leurs maris****, Vous avez toujours été présentes pour les bons conseils. Votre affection et votre soutien m'ont été d'un grand secours au long de ma vie professionnelle et personnelle. Veuillez trouver dans ce modeste travail ma reconnaissance pour tous vos efforts.*

A mes chers ***cousins****… mes chères* ***cousines****, présent dans tous mes moments importants par leurs soutiens moraux, Je vous souhaite un avenir plein de joie, de bonheur, de réussite et de sérénité. Je vous exprime à travers ce travail mes sentiments de fraternité et d'amour.*

A tous mes chers ***amis (es)*** *qui se reconnaîtront, Je ne peux trouver les mots justes et sincères pour vous exprimer mon affection et mes pensées, vous êtes pour moi des frères, sœurs et des amis sur qui je peux compter. En témoignage de l'amitié qui nous uni et des souvenirs de tous les moments que nous avons passé ensemble, je vous dédie ce travail et je vous souhaite une vie pleine de santé et de bonheur.*

A ma ***promotrice****, Permettez-moi de vous exprimer mon admiration pour vos qualités humaines et professionnelles.*

A toute la ***promotion*** *2015/2016.*

Bellouti Sabrina

Je dédie ce modeste travail à :

A celle qui représente pour moi le symbole de la bonté par excellence, la source de tendresse et l'exemple du dévouement qui n'a pas cessé de m'encourager et de prier pour moi : ma chère ***mère****, Puisse Dieu, le tout puissant, te préserver et t'accorder santé, longue vie et bonheur.*

A mon cher ***père,*** *celui qui a bercé mes rêves, Aucune dédicace ne saurait exprimer l'amour, l'estime, le dévouement et le respect que j'ai toujours eu pour toi.*

A mon très cher ***mari****, quand je t'ai connu, j'ai trouvé l'homme de ma vie, mon âme sœur et la lumière de mon chemin.*

A mes chers ***frères****, les mots ne suffisent guère pour exprimer l'attachement, l'amour et l'affection que je porte pour vous.*

A mes très chères ***tantes et leurs maris****, Vous avez toujours été présentes pour les bons conseils. Votre affection et votre soutien m'ont été d'un grand secours au long de ma vie professionnelle et personnelle. Veuillez trouver dans ce modeste travail ma reconnaissance pour tous vos efforts.*

A mes chers ***cousins****… mes chères* ***cousines****, présent dans tous mes moments importants par leurs soutiens moraux, Je vous souhaite un avenir plein de joie, de bonheur, de réussite et de sérénité. Je vous exprime à travers ce travail mes sentiments de fraternité et d'amour.*

A tous mes chers ***amis (es)*** *qui se reconnaîtront, Je ne peux trouver les mots justes et sincères pour vous exprimer mon affection et mes pensées, vous êtes pour moi des frères, sœurs et des amis sur qui je peux compter. En témoignage de l'amitié qui nous uni et des souvenirs de tous les moments que nous avons passé ensemble, je vous dédie ce travail et je vous souhaite une vie pleine de santé et de bonheur.*

A ma ***promotrice****, Permettez-moi de vous exprimer mon admiration pour vos qualités humaines et professionnelles.*

A toute la ***promotion*** *2015/2016.*

Amrouni Malika

RÈSUMÈ

Les microblogs sont devenus une source importante d'information pour l'objectif de la gestion marketing, de l'intelligence économique, y compris la réputation en ligne (e-reputation management), la détection de nouvelles et de tendance.

Les Flux microblogs sont de grande valeur en raison de leur nature directe et en temps réel. Déterminer l'information portée par un message microblog peut cependant être difficile et insignifiant à cause de l'utilisation de langue créative, de la nature hautement contextualisée et informelle de ces messages et de la longueur limitée de cette forme de communication.

Nous vous proposons une solution au problème de la détermination de l'information potée par un message microblog dans le cas de TWITER à travers un lien sémantique (jonction sémantique) : Nous ajoutons la sémantique aux tweets en identifiant automatiquement les concepts qui sont sémantiquement liés aux tweets par le bais d'une annotation dites sémantique et générer des liens vers les articles de Wikipedia correspondants. A partir de ces articles on identifie de nouveaux concepts qui par la suite peuvent être utilisés pour enrichir les tweets réduisant ainsi le besoin d'Inspection et de sélection manuel. En effet notre approche est plus générale est vise à enrichir les tweets plutôt que l'extraction d'informations de ces derniers.

MOT-CLÈS : Recherche d'information, Lien sémantique, Annotation, Microblog, Twitter, Wikipedia.

Table des matières

CHAPITRE II : La RI Sémantique Dans Les Microblogs

CHAPITRE III : Approche Proposée

Listes des figures

CHAPITRE I

CHAPITRE II

CHAPITRE III

CHAPITRE IV

Liste des tableaux

CHAPITRE I

CHAPITRE IV

INTRODUCTION GÈNÈRALE

INTRODUCTION GÈNÈRALE

Introduction

Les Microblogs sont devenus une source importante d'information pour l'objectif de certaines entreprises qui ont rapidement vu l'intérêt d'utiliser les services de microblogging. Les Flux microblogs sont de grande valeur en raison de leur nature directe et en temps réel. Ces plates-formes s'avèrent en fait très agiles dans le cadre d'une utilisation professionnelle comme dans la gestion de projet où l'utilisation la plus évidente consiste à ouvrir un compte dédié à un groupe projet. Chaque membre pouvant alors poster des liens sur des articles intéressants concernant le projet, aussi les microblogs sont utilisé dans la communication événementielle où les personnes concernées à suivre un compte dédié à un événement permet de les tenir informées de l'organisation de l'événement, ou encore dans la Communication autour de nouveau produit/service Le microblogging s'avèrera là encore un moyen très simple pour la e-reputation des produit et la gestion marketing et bien d'autre utilisation comme Communication de crise, Recrutement… .

Au-delà de ces aspects d'utilisation professionnelle et personnelle les microblogs sont un moyen de collaboration et de partage rapide et pratique. Ils sont maintenant reconnus comme un moyen important pour la diffusion de l'information

Contexte

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

Dans notre travail nous allons nous intéresser à la recherche d'information sémantique dans les microblogs plus précisément dans les messages microblogs.

Ces dernières années Twitter est devenu une des plus grandes plates-formes de microblog en ligne avec 65M des visiteurs et autour 200M des tweets par jour. Les flux de microblog sont devenus des sources inestimables de nombreux types d'analyses, y compris la réputation en ligne, la détection de nouvelles et la détection de tendance. Cependant, ces messages peuvent être insignifiant à cause de l'utilisation de langue créative, de la nature hautement contextualisée et informelle et la longueur limitée de cette forme de communication. Les approches de recherche d'information classiques peuvent se trouver obsolètes dans ce contexte, ce qui soulève la problématique lié à la recherche sémantique dans les microblogs.

Problématique

Comme dans beaucoup de scénarios de recherche dans les microblogs le but est de découvrir ce que les gens disent des produits, des marques, des personnes, et ainsi de suite. Ici, il est important de pouvoir précisément récupérer les tweets qui sont sur le sujet, y compris toute la désignation possible et d'autres variantes lexicales. Donc, au lieu de construire manuellement de longues requêtes de mot-clé, ya t'il un moyen qui capture toutes les variantes possibles de l'information porté par les messages microblogs lors d'une recherche ?

Objectif

Nous proposons une approche qui déterminer l'information portée par un message microblog dans la plate forme TWITTER en ajoutons de la sémantique aux tweets c.à.d enrichir les tweets en identifiant automatiquement les concepts clés des ces derniers. Nous prenons chaque concept pour être n'importe quel élément (article) qui a une entrée unique et sans ambiguïté dans une source de connaissance à grande échelle célèbre Wikipedia.

Notre approche proposée implique deux étapes pour l'enrichissement sémantique :

- La première étape est orientée Rappel où le but est d'obtenir une liste classée des concepts candidats toute en réduisant au maximum les pertes.

- Dans la deuxième étape, nous améliorons la précision et décidons lequel des concepts candidats à garder pour l'enrichissement des tweets.

Organisation du mémoire

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

Chapitre 1 : « De La RI classique à La RI sémantique ». Dans ce chapitre nous présentons notre domaine d'application (La recherche d'information), nous rappelons les concepts clés de ce domaine, puis nous ouvrerons le passage de La recherche d'information classique à la recherche d'information sémantique tout en décrivant cette dernière qui jouera par la suite un rôle important dans l'approche proposée.

Chapitre 2 : « La RI sémantique dans les microblogs ». Dans ce chapitre nous introduisons la recherche d'information sémantique dans les microblogs. Tout d'abord nous commençons par présenter les spécificités des microblogs, puis nous passerons en revue les différents moyens de classifications, de recherches et de détections pour l'accès à l'information dans les micros blogs. Enfin nous ferons un état de l'art sur TWITTER et la RI sémantique.

Chapitre 3 : « Approche proposée ». Dans ce chapitre nous rappelons d'abord les notions nécessaires à la compréhension de notre approche puis nous présentons notre approche ainsi que tous les travaux connexes relatifs à cette dernière.

Chapitre 4 : « Implémentation et Test ». Dans ce dernier chapitre nous implémentons notre approche puis nous testons son efficacité à travers une comparaison avec une autre approche classique, enfin nous présentons nos résultats.

Enfin, nous terminons ce mémoire par une conclusion et nous présentons nos perspectives de recherches futures.

CHAPITRE I

De La RI Classique

À

La RI Sémantique

I.1. INTRODUCTION

De nos jours, nous sommes privilégiés dans un monde riche en information, dans lequel la plupart, si ce n'est la totalité, de l'information dont nous avons besoin est au bout de nos doigts et est prête à être exploitée. Toutefois, les conséquences de ce flux d'information ont mené à la sélection de données non pertinentes, en réponse à nos requêtes d'information. L'utilisateur se voit alors dérouté et ne sait par où commencer sa quête d'information, quand la finir, s'il a eu une information correcte et la plus récente. Il ignore aussi si ce qu'il a eu représente la totalité de l'information pertinente disponible ou alors s'il existe plus d'informations pertinentes. La difficulté d'évaluer l'information désirée a donc augmenté avec la croissance du volume de l'information disponible, ce qui engendre un manque d'outils suffisants pour aller vers une exploitation maximale de l'information disponible par la prise en compte de la sémantique.

La sémantique consiste en l'étude de la signification des mots ainsi que les rapports de sens entre ces derniers (tels que l'homonymie, la synonymie, l'antonymie, l'hyperonymie, l'hyponymie). Elle constitue un des enjeux majeurs dans l'évolution des systèmes de RI en général. La prise en compte de la sémantique passe notamment par l'emploi de ressources sémantiques externes à la collection de documents initiale.

Nous pensons donc que l'avenir des systèmes de Recherche d'Information passe par la prise en compte de la sémantique du contenu des documents, permettant à un utilisateur de mieux maîtriser le flux d'information pour cibler l'information dont il a réellement besoin.

Le but de ce chapitre est d'étudier la recherche d'information dans toutes ses structures et de convergé vers une recherche d'information sémantique.

I.2. RECHERCHE D'INFORMATION

La recherche d'information est la branche de l'informatique qui s'intéresse à l'acquisition, l'organisation, le stockage, la recherche et la distribution de l'information.

La recherche d'information vise à satisfaire les besoins des utilisateurs en mettant en place un mécanisme qui va faire une correspondance entre les besoins et les documents d'une base documentaire.

I.3. SYSTEME DE RECHERCHE D'INFORMATION

Un Système de Recherche d'Information (SRI) est un système informatique constitué d'un ensemble de programmes, dont l'objectif principal est de retrouver et de sélectionner, dans une *collection de documents* préalablement enregistrée, le maximum de *documents*

pertinents répond au besoin en information exprimé par un utilisateur sous forme de *requête*. La figure ci-dessous illustre ce fonctionnement.

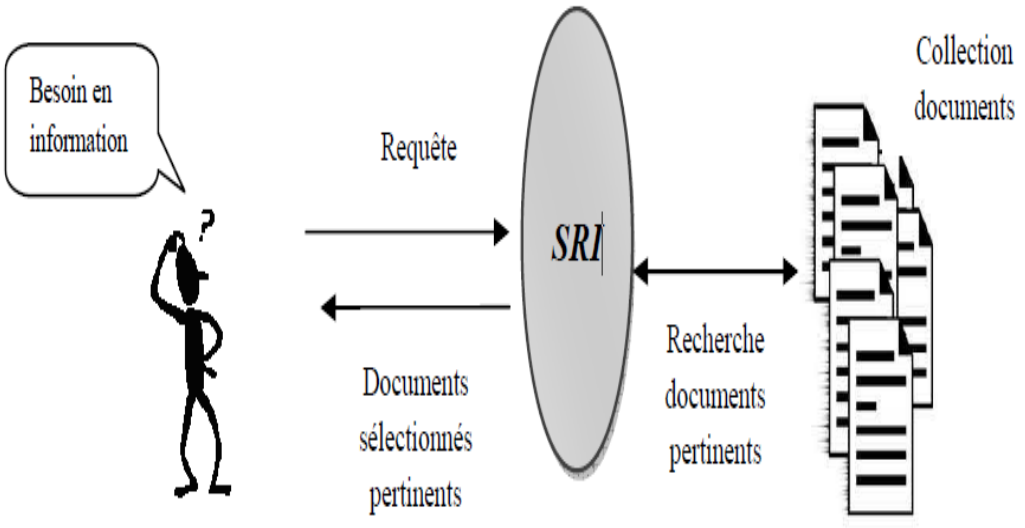


Figure I.1 : fonctionnement d'un système de recherche d'information.

Cette définition fait ressortir les quatre concepts clés suivants : collection de document, besoin en information, requête et pertinence.

- **Collection de documents**

La collection de documents (ou fond documentaire, corpus) est un l'ensemble d'informations exploitables et accessibles par un utilisateur, ou simplement, c'est l'ensemble de documents dans lequel l'utilisateur cherche une information.

- **Document**

On appelle document toute unité d'information qui peut constituer une réponse à un besoin en information/requête d'un utilisateur. Un document peut se présenter sous plusieurs formes texte, un morceau de texte, une image, une bande vidéo, etc. Il peut apparaitre dans différents langage (français, anglais, arabe,…), et peut être stocké dans de différents supports.

- **Requête :**

La requête est la représentation du besoin en information d'un utilisateur, c'est elle qui initie le processus de recherche. Elle peut être exprimée en langage naturel, booléen ou graphique, etc. la requête représente l'interface entre l'utilisateur et le système de recherche d'information (SRI).

- **Pertinence :**

Le cœur du problème de la recherche d'informations réside dans la définition d'une fonction de correspondance entre la requête et l'ensemble des documents disponibles. La pertinence en recherche d'informations peut être vue sous différents angles :

- *pertinence utilisateur;* elle représente alors la façon dont ce dernier évalue les documents retrouvés par le système de recherche d'informations en fonction de son besoin d'informations (on parle de ses jugements de pertinence). Deux utilisateurs peuvent avoir des jugements différents sur une même requête. [Abbas Nacira 14]

- *pertinence système,* elle s'exprime sous la forme d'un score obtenu automatiquement par les systèmes de recherche d'informations, en comparant les représentations des documents et celles des requêtes suivant les méthodes définies par le modèle de recherche d'informations utilisé.

La notion de pertinence est difficile à automatiser, car elle est fortement subjective, c'est à dire dépendante de l'utilisateur. Le but du SRI est alors de faire correspondre au mieux la pertinence système avec la pertinence utilisateur

I.4. PROCESSUS DE LA RI

Pour répondre aux besoins en information de l'utilisateur un SRI met en œuvre un certains nombre de processus pour réaliser la mise en correspondance des informations contenues dans un fonds documentaire disponibles d'une part, et les besoins en information des utilisateurs d'autre part. Ces besoins sont traduits de façon structurée par l'utilisateur sous forme de requêtes.

Le système de recherche d'information (SRI) intègre trois fonctions principales représentées schématiquement par le processus en U de recherche d'information :

-L'indexation : consiste à extraire et à représenter le contenu des documents de manière interne sous forme d'index. Cette structure d'index permet de retrouver rapidement les documents contenant les mots clés de la requête.

-L'appariement requête-document : vise à apparier les documents et la requête de l'utilisateur en comparant leurs descripteurs respectifs. Pour cela, le système calcule la pertinence de chaque document vis-à-vis de la requête utilisateur selon une mesure de correspondance du modèle de RI, et retourne la liste des résultats à l'utilisateur.

-La reformulation de la requête : La reformulation de la requête consiste à enrichir la requête utilisateur en ajoutant des termes permettant au mieux exprimer son besoin.

De manière générale, la recherche dans un SRI consiste à comparer la représentation interne de la requête aux représentations internes des documents de la collection.

- ✓ Formulation de la requête par l'utilisateur, dans un langage de requêtes naturel, à base de mots clés ou le langage booléen.
- ✓ La requête formulée sera transformée en une représentation interne équivalente, lors d'un processus d'interprétation.
- ✓ Un processus similaire, dit indexation, permet de construire la représentation interne des documents de la base documentaire.
- ✓ Le processus de recherche met en correspondance et calcule le degré d'appariement des représentations internes des documents et de la requête.
- ✓ Les documents qui correspondent au mieux à la requête, ou documents dits pertinents, sont alors retournés à l'utilisateur.
- ✓ Afin d'améliorer les résultats de la recherche, le système peut être doté d'un mécanisme d'amélioration et de raffinement de la requête par reformulation.
- ✓ Le fonctionnement général d'un SRI est donné au travers du processus en U [Belkin et al, 92], présenté en figure I.2.

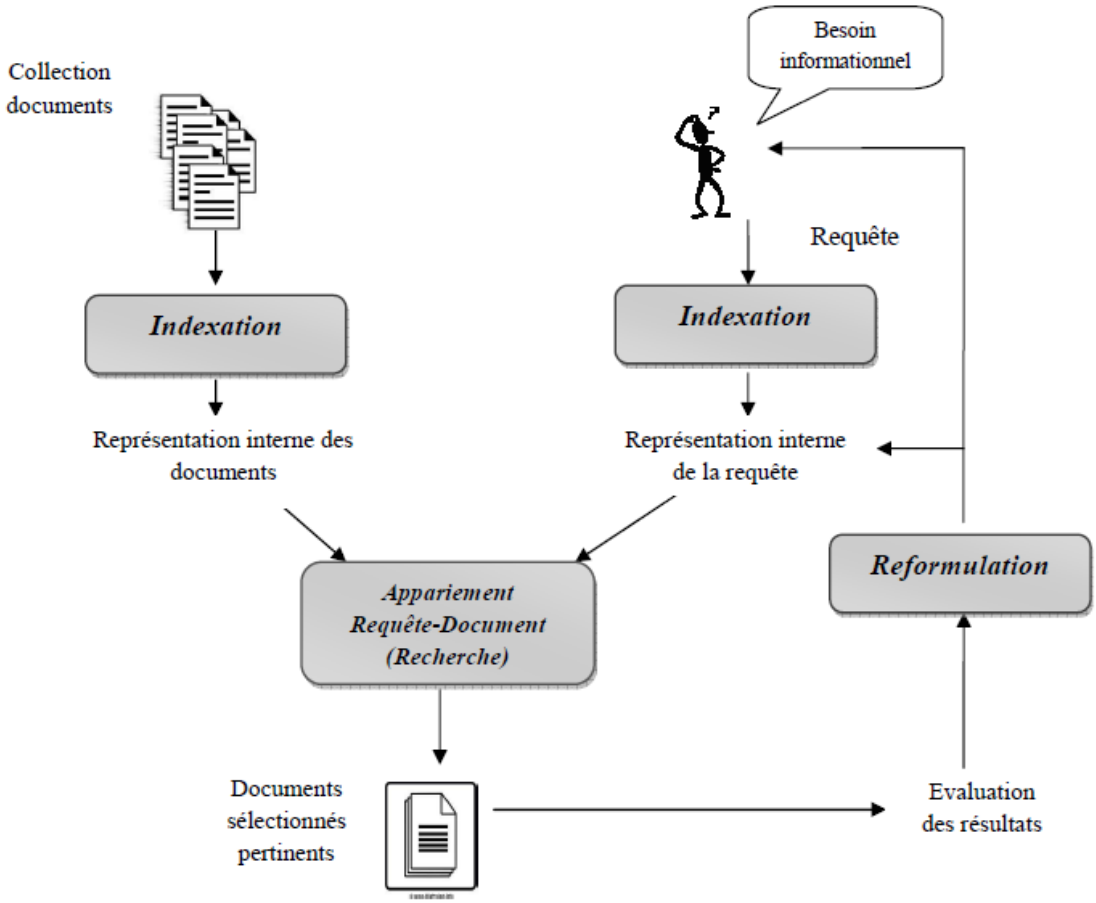


Figure I.2 : Processus en U de Recherche d'Information

I.4.1. Processus d'indexation

Pour que le coût de la recherche soit acceptable, il convient d'effectuer une étape primordiale sur la collection de documents. Cette étape consiste à analyser chaque document de la collection afin de créer un ensemble de mots clés : on parle de l'étape d'indexation. Ces mots clés seront plus facilement exploitables par le système lors du processus ultérieur de recherche. L'indexation permet ainsi de créer une représentation interne de contenu des documents dans le système. Son objectif est de trouvé les concepts les plus importants du document (ou de la requête), qui formeront *le descripteur du document* qui peuvent être :

A la fin de cette étape, les documents sont représentés dans des fichiers index qui stockent la cartographie des couples terme-document en y associant un poids.

L'indexation peut être :

- Manuelle : chaque document est analysé par un spécialiste du domaine ou par un documentaliste ;
- Automatique : le processus d'indexation est entièrement informatisé ;
- Semi-automatique : le choix final revient au spécialiste ou le documentaliste, qui intervient, après l'indexation automatique, pour finalisé le choix des termes significatifs.

L'indexation manuelle, permet d'assurer une meilleure pertinence dans les réponses apportées par le SRI. Elle présente toute fois plusieurs inconvénients : deux indexeurs différents peuvent présenter des termes différents pour caractériser un même document, et un indexeur à deux moments différents peut présenter deux termes distincts pour représenter le même concept. De plus, le temps nécessaire à sa réalisation est très important.

L'indexation semi-automatique [BAL 95], appelée aussi indexation supervisée, est une combinaison entre l'indexation manuelle et l'indexation automatique. Dans ce cas, les indexeurs utilisent *un thesaurus* ou une base terminologique, qui est une liste organisée de descripteurs (mots-clés) obéissant à des règles terminologiques propres et reliés entre eux par des relations sémantiques.

En fin, ***l'indexation automatique*** [MAR 60], est la plus étudiée en RI. Elle regroupe un ensemble de traitements automatisés sur un document en se basant sur la construction des descripteurs de manière automatique et rapide. L'indexation automatique repose sur des approches plus répandue qui sont : l'extraction automatique des mots des documents (analyse lexicale), l'élimination des mots vides en utilisant un anti-dictionnaire, lemmatisation (radicalisation ou normalisation), le repérage de groupes de mots, la pondération des mots avant de créer l'index.

A la fin de cette étape, les documents sont représentés dans des fichiers index qui stockent la cartographie des couples terme-document en y associant un poids.

☞ Analyse lexicale

L'analyse lexicale permet de convertir le texte d'un document en un ensemble de termes. Un terme est une unité lexicale ou un radicale (mots simples ou composés) [FOX 92]. Elle permet de reconnaitre les espaces de séparation des mots, les chiffres, les ponctuations, etc

☞ Élimination des mots vides

Un des problèmes majeurs de l'indexation est d'arrivé à extraire les termes significatif tout en évitant les mots vides (pronoms personnels, propositions, etc.). Ces mots vides peuvent aussi être des mots athématiques (les mots qui peuvent se retrouvé dans n'importe quel document parce qu'ils exposent le sujet mais ne le traite pas, comme par exemple *contenir, appartenir*, etc.). On distingue deux techniques pour éliminer les mots vides :

- L'utilisation d'une liste de mots vides (aussi appelée anti-dictionnaire),
- l'élimination des mots dépassant un certain nombre d'occurrences dans la collection.

Même si l'élimination des mots vides a l'avantage évident de réduire le nombre de termes d'indexation, elle peut cependant réduire le taux de rappel, c'est-à-dire la proportion de documents pertinents renvoyés par le système par rapport à l'ensemble des documents pertinents. Par exemple, en éliminant le mot *a* de *vitamine a*, les documents pertinents qui contiennent ce dernier terme ne sont pas retournés par le SRI.

☞ Lemmatisation

Un mot donné peut avoir différentes formes dans un texte, mais leur sens reste le même ou très similaire. On peut par exemple citer économie, économique, économétrie, etc. Il n'est pas forcement nécessaire d'indexer individuellement tous ces mots alors qu'un seul suffirait à représenter le concept véhiculé. Pour résoudre le problème, une substitution des termes par leur racine, ou lemme, est utilisée.

Cette phase de passage à la forme canonique n'est pas obligatoire. Elle présente le principal avantage d'indexer par exemple le mot *camions* et le mot *camion* de la même façon, ce qui évite à l'utilisateur de devoir entrer les formes de pluriel des noms ou les formes conjuguées des verbes lors de sa recherche. Cependant, dans certains cas, le passage à la forme canonique supprime la sémantique originale du mot. Par exemple, selon le mode de lemmatisation, la forme conjuguées *portera* du verbe *porter* sera indexée sous *porte*, de la même façon que le mot *portes*. Ainsi, lorsque l'utilisateur formulera une requête avec verbe *porter*, il aura très certainement, parmi la liste des documents sélectionnés, des documents non pertinents relatifs au nom *porte*. Si la lemmatisation a pour but d'augmenter le rappel, la précision (c'est-à-dire la proportion de documents pertinents par rapport au nombre de documents renvoyées par le système) en fait souvent les frais.

Pour résoudre ce problème, C Crouch [CRO 02] propose une méthode en deux temps, dont les résultats s'avèrent encourageants :

- une première recherche est effectuée, en utilisant une lemmatisation des mots ;
- les documents sont ensuite réordonnés en fonction de la présence des termes non-lemmatisés de la requête dans leur contenu.

☞ **Pondération des termes**

La pondération est l'une des fonctions fondamentales en RI. Elle est la clé de voûte de la majorité des méthodes et approches en RI proposés depuis les années 1960. Le poids d'un terme dans un document traduit l'importance de ce terme dans ce document. Cette importance est souvent calculée sur la base d'aspects statistiques.

La plupart des techniques de pondération sont basées sur la combinaison de deux facteurs : une pondération locale et une pondération globale.

- La pondération locale (*term frequency, notée Tf*) : cette mesure est propositionnelle à la fréquence du terme dans le document.

La pondération globale (*inverse document frequency, notée Idf*) : Cette mesure est propositionnel à la fréquence d'un terme (t) dans toute la collection (d). En effet, un terme qui apparaît souvent dans la base documentaire ne doit pas avoir le même impact qu'un terme moins fréquent. Il est généralement exprimé comme suit :

$$\mathbf{Log\ (N/d_j)} \qquad (1.1)$$

Où :

- **N *:*** est le nombre total de documents da la base documentaire.
- **D_j *:*** est le nombre de documents contenant le terme.

- La pondération *tf*idf* : Une nouvelle approche de pondération pour 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 :

$$\mathbf{TFIDF(t,d) = TF(t,d\ ) * IDF(t)} \qquad (1.2)$$

I.4.2. L'appariement requête-document

Une fois les documents sont représentés sous forme interne d'index, il est possible de rechercher ceux qui répondent le mieux à une requête d'un utilisateur grâce à la relation d'appariement.

Dans ce processus, le système « décide » quels sont les documents qui correspondent à la requête de l'utilisateur, et permet de classer les documents par ordre de pertinence pour la requête.

Le système calcule un score de correspondance entre la représentation de chaque document et celle de la requête. Ce score exprime un degré de pertinence système. Ce dernier est calculée à partir d'une fonction de similarité appelée RSV (Q,D) ou Q est une requête et D un document.

I.4.3. Reformulation de requêtes

La satisfaction de l'utilisateur concernant les documents retournés par le système, n'est pas toujours acquise. Certains systèmes permettent aux utilisateurs de marquer parmi les documents résultats ceux qu'ils jugent pertinents ou non pertinents. Ces jugements sont alors pris en compte pour définir une nouvelle requête, il s'agit du processus de reformulation. Cette étape a pour objectif d'améliorer les performances du SRI par reformulation de la requête initiale en y ajoutant de nouveaux termes significatifs et/ou par ré-estimation de leur poids.

La reformulation de la requête est alors considérée comme un processus ayant pour objectif de générer une nouvelle requête plus ciblée permettant d'obtenir des résultats de recherche plus pertinents que ceux obtenus par la requête initialement formulée par l'utilisateur.

I.5. MODELE DE RI

Un modèle de RI a pour rôle de fournir une formalisation du processus de recherche d'information. Il doit accomplir plusieurs rôles dans le plus important est de fournir un cadre théorique pour la modélisation de la mesure de pertinence. Cette mesure notée par RSV (Retrieval Status Value), détermine, lors du processus d'appariement-requête, le degré de ressemblance entre les représentations du document et les requêtes. Ces modèles peuvent être divisés en deux catégories :

1. les modèles dits « exacts » qui ne retournent que des documents répondant exactement à la requête (modèle booléen). ou
2. les modèles dits « partiels » (probabiliste, vectoriel…) qui retournent des documents répondant à tout ou partie de la requête.

Les modèles booléens sont les premiers modèle utilisés en RI, ils sont inspirés de la logique booléenne et de la théorie des ensembles pour modéliser l'appariement document-

requête. On distingue trois variations principales : le modèle booléen classique, le modèle booléen étendu et le modèle booléen flou.

Dans les modèles vectoriels, la représentation des documents et requêtes est réalisée par des vecteurs de termes pondérés dans l'espace vectoriel multidimensionnel des termes d'index. Ils englobent le modèle vectoriel généralisé, le modèle LSI (Latent Semantic Indexing) et le modèle connexionniste.

Les modèles probabilistes sont basés sur les probabilités d'appartenance des termes de la requête aux documents de la collection. Ils ont été introduits pour modéliser la notion de pertinence. Ils composent de modèle probabiliste général, de modèle de réseau inférentiel (Document Network), et de modèle de langue.

Ces modèles utilisent une *valeur réelle (degré de pertinence système)* pour rendre compte du degré de l'appariement entre la requête et une unité documentaire. Il est à noter que cette pertinence système est une valeur calculée et qu'elle peut être différente de la pertinence réelle qui découle du jugement de pertinence réalisé par l'utilisateur.

MODELES		
EXACTS	**PARTIELS**	
Booléen	**Vectoriel**	**Probabilistes**
Modèle booléen classique	Modèle vectoriel généralisé	Modèle probabiliste général
Modèle booléen pondéré (étendu)	Modèle LSI (Latent semantic Indexing)	Modèle de réseaux de neurones
Modèle booléen flou	Modèle connexionniste.	Modèle de langue

Tableau I.1 : Les modèles de la RI

I.5.1. Modèles booléens

Le modèle booléen a été introduit en 1983 par Salton et McGill. Il s'est imposé grâce à la simplicité et à la rapidité de sa mise en œuvre [Salton 83]. L'interface d'interrogation de la plupart des moteurs de recherche (Google, Alta Vista) est basée sur les principes de ce modèle.

Ce modèle est basé sur la théorie des ensembles et l'algèbre booléenne. Un document est représenté par l'ensemble de ses termes descriptifs (non pondérés) qui constitue l'index

du document, Un exemple de représentation d'un document est comme suit : **$D=t_1$ ET t_2 ET $t_3 \ldots t_n$.** Une requête est une expression booléenne composée des mots clés reliés par des opérateurs logiques **ET (∧), OU (∨) et NON (¬).** Un exemple de représentation d'une requête est comme suit **(t_1 ET t_2) OU (t_3 ET t_4)…..**
La mesure de pertinence document-requête est calculée selon une fonction de correspondance qui est basée sur l'hypothèse de présence/absence des termes de la requête dans le document et vérifie si l'index de chaque document $\mathbf{D_i}$ implique l'expression logique de la requête **q**. Le résultat de cette fonction est donc binaire est décrit comme suit :

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

RSV (d, $q_1 \wedge q_2$) = 1 **si RSV (d, q_1) = 1 et RSV (d, q_2) = 1 ; 0 sinon**

RSV (d, $q_1 \vee q_2$) = 1 **si RSV (d, q_1) = 1 ou RSV (d, q_2) = 1 ; 0 sinon**

RSV (d, ¬q) = 1 **si RSV (d, q) = 0 ; 1 sinon**

Ce modèle présente plusieurs avantages tels que simplicité de mise en œuvre et la clarté de l'expression de la requête grâce à des opérateurs logiques.

L'inconvénient majeur de ce modèle, est que les documents pertinents dont la représentation ne correspond qu'approximativement à la requête ne sont pas sélectionnés, et que tous les termes ont la même importance. Pour remédier à ces inconvénients, SALTON a proposé des extensions de ce modèle comme le modèle booléen étendu et le modèle booléen basé sur les ensembles flous [Wartik Steven 92]. Ces modèles sont basés sur le principe des pondérations. Il tient compte de l'importance des termes dans la représentation des documents et dans la requête, et ce, en affectant des poids à chaque terme du document et de la requête.

I.5.2. Le modèle vectoriel

Le modèle vectoriel est un modèle algébrique, introduit par Salton [Salton 75]. Il représente chaque document, ainsi que la requête, par un vecteur et calcule un coefficient de similarité (**RSV**) entre chaque document et la requête. Ce coefficient de similarité correspond, par exemple, au cosinus des angles entre le vecteur de la requête et le vecteur d'un document, afin de trouver les documents dont le vecteur de représentation est le plus colinéaire avec le vecteur de la requête.
La pertinence entre un document et une requête est plus grande, si le vecteur d'un document et la requête sont plus proches.

Dans ce modèle chaque document est représenté par un vecteur de dimension n, comme suit :

$$\mathbf{d_i = (d_{i1}, d_{i2}, \ldots, d_{in})} \quad \text{pour } i = 1, 2, \ldots, m.$$

Avec

- $\mathbf{d_{ij}}$ est le poids du terme $\mathbf{t_j}$ dans le document $\mathbf{D_i}$
- **m** est le nombre de documents dans la collection,
- **n** est le nombre de termes d'indexation.

Et chaque requête est représentée par un vecteur de mots-clés comme suit :

$$\mathbf{Q = (q_1 , q_2 , \dots , q_n)}$$

Où :

- $\mathbf{q_j}$ est le poids de terme $\mathbf{t_j}$ dans la requête **Q.** Ce poids peut être soit une forme de **tf*idf,** soit un poids attribué manuellement par l'utilisateur.

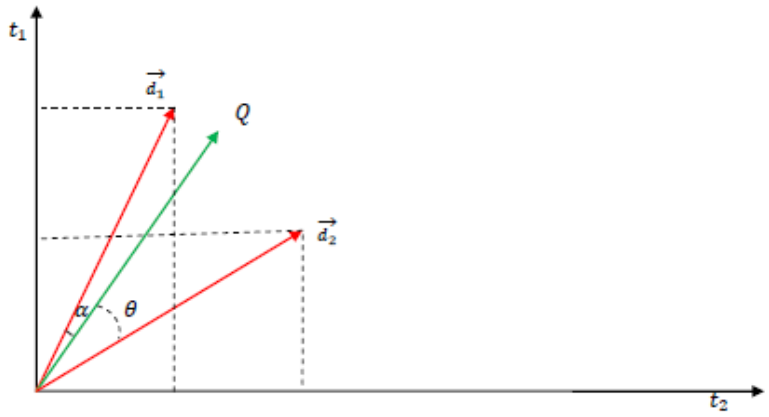


Figure I.3 : Exemple de représentation des documents dans un espace de deux termes

La pertinence du document $\mathbf{d_i}$ pour la requête **Q** est mesurée par la similarité entre leurs vecteurs correspondants. Les mesures de similarité utilisées sont :

☞ **Le produit scalaire**

$$\text{RSV (dj , Q)} = \sum_{j=1}^{n} q_j * d_{ij} \qquad (1.4)$$

☞ **la mesure de Jacard**

$$\text{RSV (dj , Q)} = \frac{\sum_{j=1}^{n} qj*dij}{\sum_{j=1}^{n} q_j{}^2 + \sum_{j=1}^{n} d_{ij}^2 - \sum_{j=1}^{n} q_j * d_{ij}} \qquad (1.5)$$

☞ **La mesure du cosinus**

Est la plus répandue des mesures de similarité utilisées pour ce modèle de recherche. Elle mesure l'angle entre les vecteurs **Q** et $\mathbf{d_i}$. Elle est équivalente au produit scalaire normalisé des vecteurs. La mesure de cosinus est donnée par :

$$\text{RSV (dj, Q)} = \frac{\sum_{j=1}^{n} qj * dj}{\sqrt{\sum_{j=1}^{n} {q_j}^2} * \sqrt{\sum_{j=1}^{n} d_j^2}} \qquad (1.6)$$

Où :

- $\mathbf{d_{ij}}$ = poids du terme $\mathbf{T_j}$ dans le document $\mathbf{D_i}$.
- $\mathbf{q_j}$ = poids du terme $\mathbf{T_j}$ dans la requête **Q**.

I.5.3. Le modèle probabiliste

Le premier modèle probabiliste a été proposé par Maron et Kuhns [Mar 60] au début des années 1960. Ce modèle est basé sur la théorie des probabilistes, il consiste à calculer la pertinence d'un document en fonction de pertinences connues pour d'autres documents.
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.

Dans ce modèle, la similarité entre un document $\boldsymbol{d_i}$ et une requête ***Q***, est mesurée par le rapport entre la probabilité qu'un document soit pertinent pour une requête ***Q*** (notée ***p*** (P/d_i) et la probabilité qu'il ne soit pas pertinent à ***Q*** notée ***p*** (NP/d_i).

Etant donné une requête utilisateur notée **Q** et un document **d**.

On distingue deux classes de documents pour une requête q :

- **R** l'ensemble des documents pertinents et
- **NR** l'ensemble des documents non pertinents.

On définit **P(R|D)** comme la probabilité que le document **D** appartienne à l'ensemble des documents pertinents, et
P(NR|D) comme 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 :

$$\text{RSV (d, Q)} = \frac{p(R/d)}{p(NR/d)} \qquad (1.7)$$

Ce qui donne, d'après le théorème de Bayes et après simplification :

$$\text{RSV}(d, Q) = \frac{p(R/d)}{p(NR/d)} \approx \frac{p(d/R)}{p(d/NR)} \quad (1.8)$$

Tel que :

- **P(d/R)** est la probabilité que le document appartienne à l'ensemble **R** des documents pertinents
- **P(d/NR))** est la probabilité que le document appartienne à l'ensemble **NR** des documents non pertinents).

Plusieurs solutions ont été proposées pour représenter le document **D**. Parmi elles, on cite le modèle BIR (Binary Independance Retrieval) [Rijsbergen, 79], qui repose sur l'indépendance des termes dans les documents. Ainsi chaque document $\mathbf{d_i}$ de la collection est représenté par un ensemble d'événements qui dénotent la présence ou l'absence d'un terme $\mathbf{t_j}$ dans un document.

Formellement :

$\mathbf{d_i = (t_1 = x_1, t_2 = x_2, \ldots, t_j = x_j, t_n = x_n)}$ où $x_j = \begin{cases} \mathbf{1} & \boldsymbol{si\ t_j\ est\ présent dans\ d_i} \\ \mathbf{0} & \boldsymbol{sinon} \end{cases}$

En supposant que ces événements sont indépendants, les probabilités de pertinence et de non-pertinence P (d/R) et P(d/NR) sont calculées comme suit :

$$P(d_i/R) = \prod_{j=1}^{n} p(t_j = x_j/R) \quad (1.9)$$

$$P(d_i/NR) = \prod_{j=1}^{n} p(t_j = x_j/NR) \quad (1.10)$$

L'application de la distribution de la loi Bernoulli sur les probabilités $\mathbf{P(d_i/R)}$ et $\mathbf{P(d_i/NR)}$, permet d'obtenir les résultats suivants :

$$P(d_i/R) = \prod_{j=1}^{n} p(t_j = x_j/R) = \prod_{j=1}^{n} p(t_j = 1/R)^{x_j} * p(t_j = 0/R)^{1-x_j} \quad (1.11)$$

$$P(d_i/NR) = \prod_{j=1}^{n} p(t_j = x_j/NR) = \prod_{j=1}^{n} p(t_j = 1/NR)^{x_j} * p(t_j = 0/NR)^{1-x_j} \quad (1.12)$$

Notons que :

- p_i = P (t_i=1/R) entraine 1-p_i = P (t_i=0/R), tel que : p_i est la probabilité qu'un terme $\mathbf{t_j}$ soit présent dans l'ensemble des documents pertinents.

- q_i = P (t_i=1/NR) entraine 1-q_i=P(t_i=0/NR), tel que : q_i est la probabilité qu'un terme $\mathbf{t_j}$ soit présent dans l'ensembles des documents non pertinents.

Alors, les probabilités de pertinence et de non-pertinence P(d/R) et P(d/NR) sont calculées comme suit :

$$P(d_i/R) = \prod_{j=1}^{n} p(t_j = x_j/R) = \prod_{j=1}^{n} p_i^{x_j}(1-p_i)^{1-x_j} \qquad (1.13)$$

$$P(d_i/NR) = \prod_{j=1}^{n} p(t_j = x_j/NR) = \prod_{j=1}^{n} q_i^{x_j}(1-q_i)^{1-x_j} \qquad (1.14)$$

Après quelques transformations, la fonction RSV (d, Q) peut s'écrire comme suit :

$$\text{RSV (d, Q)} = \frac{\text{p(d/R)}}{\text{p(d/NR)}} = \frac{\prod_{j=1}^{n} p_i^{x_j}(1-p_i)^{1-x_j}}{\prod_{j=1}^{n} q_i^{x_j}(1-q_i)^{1-x_j}} \qquad (1.15)$$

$$\text{RSV (d, Q)} = \sum_{\text{i=1}} \log \frac{\text{p}_\text{i}\,(1-\text{q}_\text{i})}{\text{q}_\text{i}\,(1-\text{p}_\text{i})} \qquad (1.16)$$

on suppose connus l'ensemble **R** des documents pertinents et l'ensemble **NR** des documents non pertinents, alors on peut estimer les probabilités $\mathbf{p_i}$ et $\mathbf{q_i}$, en utilisant les proportions définies en Tableau I.2, comme suit :

$$p_i = \frac{r_i}{n} \qquad et \qquad q_i = \frac{R_{i\,-r_i}}{N-n}$$

	Pertinent	Non pertinent	Total
Terme t_i présent	r_i	$n_i - r_i$	N
Terme t_i non présent	$R_i - r_i$	$N - _i - (R_i - r_i)$	N – n
Total	R_i	$N - R_i$	N

Tableau I.2 : Distribution de probabilités de pertinence des termes d'un corpus d'apprentissage

Avec :

- R_i : nombre de documents contenant le terme t_i.
- N: nombre de documents dans le corpus.
- r_i: nombre de documents pertinents contenant t_i
- n : nombre de documents pertinents contenant le terme t_i

Ainsi la fonction de similarité RSV se réduit à :

$$\text{RSV}(d, Q) = \sum_{i=1} \log \frac{r_i\,(N - R_i - n - r_i)}{(n - r_i)\,(R_i - r_i)} \qquad (1.17)$$

I.6. EVALUATION DES SRI

Un système de recherche d'information doit satisfaire un besoin d'information d'un utilisateur c'est à dire de trouver seulement les documents pertinents. La Pertinence est la notion centrale dans la RI car toutes les évaluations s'articulent autour de cette notion qui est définit comme étant la correspondance entre un document et une requête, 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 correspond à celles que l'utilisateur espère, mieux est le système.

I.6.1. Collections de test

Une collection de test est utilisée en RI pour évaluer les stratégies de recherche afin d'assurer l'efficacité des SRI. Plus particulièrement, une collection de test doit traduire la

subjectivité de pertinence des utilisateurs. Elle est composée d'un ensemble de documents, d'un ensemble de requêtes, ainsi que des jugements de pertinence établis manuellement.
La collection test est constituée lorsque l'on dispose, pour une collection de documents donnée, d'un ensemble de questions-tests avec leurs « réponses idéales » associées. Les réponses idéales vont permettre d'évaluer la qualité des réponses fournies par des systèmes documentaires à évaluer.

Différentes collections de test sont utilisées en recherche d'information pour évaluer les performances d'un SRI, la plus connu est la collection TREC (voir annexe) adopté dans les compagnes TREC. Son but est d'examiner les premiers documents pertinents retournés par un système, puis différentes mesures d'évaluation sont utilisées pour calculer la performance du système à partir de ces résultats de recherche.

I.6.2. Mesure d'évaluation

Cegroupedemesuresprendencompteuniquementlenombrededocumentspertinents 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.

D'une façon générale, tout SRI a deux objectifs principaux : retrouver tous les document pertinants, et rejeter tous les documents non-pertinents. Plusieurs mesures standards en RI ont été proposées pour évaluer les performances des SRI [MOH 08], [JAC 08]. les deux princpiaux groupes de mesures permettant d'évaluer un SRI sont les mesures non-ordonnées et les mesures ordonnées

➲ 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 le rappel et la précision

Rappel et précision :

La *précision*, notée **P**, mesure le ratio entre le nombre de documents pertinents trouvés et le nombre total de documents trouvés (car on cherche à trouver uniquement des documents pertinents et non du bruit), ce qui donne un degré de solidité du classificateur. Une précision égale à 1 signifie que le système n'a retrouvé que des documents pertinents. Elle est exprimée par :

$$P = \frac{|\text{Documents pertinents trouvés (Ra)}|}{|\text{Documents trouvés (A)}|} \qquad (1.18)$$

Le *rappel* noté **Rp** mesure le ratio entre le nombre de documents pertinents trouvés et le nombre de documents pertinents trouvés présents dans la base (car on cherche à trouver tous les documents pertinents et non du silence), ce qui donne un degré d'exhaustivité.

Toute analyse cherche donc à limiter le bruit et à éviter le silence. Or des tests de robustesse doivent être réalisés pour faire ce calcul et démontrer la qualité du travail du classificateur. Un rappel égal à 1 signifie que tous les documents pertinents ont été retrouvés. Il est exprimée par :

$$\mathrm{Rp} = \frac{|\text{Documents pertinents trouvés (Ra)}|}{|\text{Documents pertinents (R)}|} \qquad (1.19)$$

La *F-mesure,* est une mesure qui combine la précision et le rappel, nommée F-mesure ou F-score introduite dans [Rijsbergen, 79] est définie par :

$$\mathrm{F-mesure} = \frac{2*\mathrm{P}*\mathrm{Rp}}{\mathrm{P}+\mathrm{Rp}} \qquad (1.20)$$

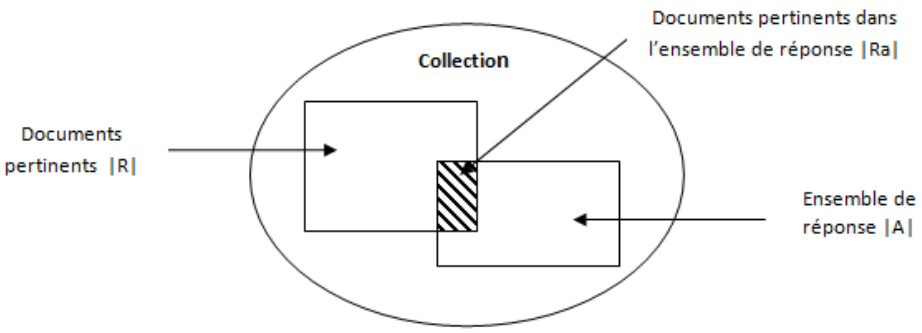


Figure I.4 : Précision et Rappel

➲ Mesures ordonnées

Ce groupe de mesures prend en compte 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.

$$\mathrm{P@X(q)} = \mathrm{P_t/X} \qquad (1.21)$$

Où Pt est le nombre de documents pertinents.

R-précision:

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

La Précision Moyenne (Average precision-AP) :

C'est la moyenne des valeurs de précisions après chaque document pertinent, elle se calcule comme suit :

$$AP_q = 1 / R \sum_{i=1}^{N} p(i) \times R(i) \qquad (1.22)$$

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
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.

$$MAP = \sum_{q \in Q} AP_q \; / \; |Q| \qquad (1.23)$$

I.7 LES DIFFICULTES DE LA RI CLASSIQUE

Dans la majorité des systèmes de recherches d'information classique, l'indexation d'un document ou d'une requête est représentée par des mots clés. Cette représentation est imprécise ne prend pas en considération la sémantique des mots, ce qu'implique deux majeur de problèmes : l'ambiguïté des mots et leurs disparité lors de la recherche. [Abbas Nacira 14] [Azzouz Wassila 13]. Ce que signifie que le système retourne à l'utilisateur des documents non pertinents.

-L'ambiguïté des mots : fait référence à un mot ayant plusieurs sens. Ce problème implique qu'un système peut retourner des documents non pertinents contenant les mêmes mots de la requête mais avec des sens différents. Ce qu'engendre un bruit. [Krovetz et al 92].

-La disparité des mots : se réfère à des mots lexicalement différents mais portant un même sens. Ce problème implique qu'un document ne partage pas de termes avec la requête, peut ne pas être retournée même s'il est pertinent. [Egozi et al, 11].

Pour pallier ces problèmes, plusieurs approches d'indexation ont été proposés dans l'objectif principal est de représenter les documents et les requêtes par des concepts (sens des mots) qui sont par nature non ambigus.

I.7.1. Approche de la Recherche d'Information Sémantique

Généralement, La recherche sémantique est une présentation de réponses à une recherche, d'une façon qui essaie de comprendre la requête de l'utilisateur et l'intention qui la sous-tend. Pour y parvenir, le moteur de recherche doit comprendre le langage naturel de la même façon que l'être humain, ce qui signifie comprendre non seulement les mots mais aussi leur signification contextuelle. Pour cela, plusieurs travaux tentant d'incorporer l'information sémantique dans le processus de RI. Au sein de ces travaux, deux grandes approches peuvent être distinguées [Mihalcea et al., 00] [Biemann, 05] : l'indexation conceptuelle et l'indexation sémantique. En effet, on parle d'indexation sémantique quand il s'agit d'utiliser le sens des mots (mot-sens ou word-sens) pour indexer les documents (telle que utilisée par Sanderson [Sanderson, 94] et Krovetz et Croft [Krovetz et al., 92]). Partant de cette définition, l'indexation conceptuelle peut être vue comme une généralisation de l'indexation sémantique, dans la mesure où les concepts aussi véhiculent des sens. Cependant, nous avons décidé de les séparer pour deux raisons :

(1) la première est due au fait que l'indexation sémantique en RI se base historiquement sur les techniques de désambiguïsation (WSD) pour affecter un sens à un mot, alors que l'indexation conceptuelle se base sur des méthodes d'identification de concepts dans un corpus textuel (appariement de concept).

(2) la deuxième raison est que, dans l'indexation conceptuelle, la structure conceptuelle utilisée rend possible une extension de la représentation des documents (ou requêtes) via les différentes relations sémantiques qu'elle procure. Or cela n'est pas possible avec l'indexation sémantique, où l'extension de la représentation se limite souvent à l'utilisation de la synonymie.

I.7.1.1. L'indexation sémantique ou terminologique

L'indexation sémantique est une approche d'indexation basée sur le sens des mots [BAZIZ, 2005]. Elle s'appuie sur des algorithmes de désambiguïsation de mots (WSD) pour indexer les documents et les requêtes avec le sens des mots (mots-sens) plutôt qu'avec des mots simples. Une manière d'indexer serait, par exemple, d'associer aux mots extraits, des mots du contexte qui aident à déterminer leur sens. D'autres approches de désambiguïsation plus élaborées, utilisent des représentations hiérarchiques pour calculer la distance sémantique ou similarité sémantique entre les mots à comparer.

L'indexation sémantique se base essentiellement sur le processus de « désambigüisation » qui permet, dans l'idéal, de ramener tous les mots à leur sens originel. Deux scénarios antagonistes existent [BAZIZ, 2005], [DESMONTILS, JACQUIN, & MORIN, 2002] :

Il faudrait que l'outil de désambiguïsation soit capable de désambiguïser de façon exacte une collection de documents et qu'un SRI soit capable de traiter une telle collection, il

est facile de croire que l'efficacité d'un tel SRI serait améliorée grâce à cette représentation plus sophistiquée.

La désambiguïsation est parfois peu susceptible d'être utile. Un cas usuel est celui où l'utilisateur utilise dans sa requête plusieurs termes sémantiquement proches. Par conséquent, la désambiguïsation de ce document n'est pas nécessaire étant donné la nature « auto-désambiguïsante » de la requête. Nous présentons quelques méthodes proposées pour l'indexation sémantique comme la méthode de Voorhees et la méthode de Krovetz & Croft (voir l'annexe)

I.7.2.2. L'indexation conceptuelle

L'indexation conceptuelle consiste à indexer un document, non plus avec les termes des documents, mais avec les concepts d'une base de connaissances [Stairmand et Black, 97]. Dans ces approches, il faut une « listes » de concepts cibles (qui exprime le sens des termes possibles) pour pouvoir transformer le terme en concept. Les concepts sont tirés d'un vocabulaire contrôlé : les dictionnaires de synonymes, les ontologies, les thésaurus, les taxonomies, etc. Parmi les approches qui s'inscrivent dans cet état d'esprit : le modèle OntoSeek et le modèle DocCore (voir l'annexe). [Abbas Nacira 2014]

I.8. RESSOURCES SEMANTIQUE

Les méthodes de recherche d'information sémantique visent à s'affranchir les problèmes de la recherche d'information classique via le passage au niveau conceptuel. Elles reposent souvent sur l'utilisation d'une ressource sémantique externe. En 1968, Salton constate déjà que l'utilisation de ces ressources permet d'améliorer les performances, à condition que les termes utilisés pour l'enrichissement soient validés manuellement par l'utilisateur. Il constate également que l'expansion automatique utilisant l'ensemble des termes possibles, dégrade ces performances [Salton, 1968].

Différents types de ressources sémantiques peuvent être distingués parmi lesquels se trouvent les thesaurus, ainsi que les ontologies. Les ressources sémantiques peuvent se distinguer selon l'étendue des connaissances qu'elles comportent.

I.8.1. Taxonomie

Une taxonomie est un vocabulaire contrôlé organisé sous une forme hiérarchique simple. Les liens hiérarchiques dans une taxonomie correspondent à des liens de spécialisation affinant le sens d'un terme. A la différence du thesaurus qui permet de parcourir la hiérarchie de manière connexe permettant ainsi de restreindre ou de spécialiser le champ des connaissances, cela n'est pas possible avec une taxonomie.

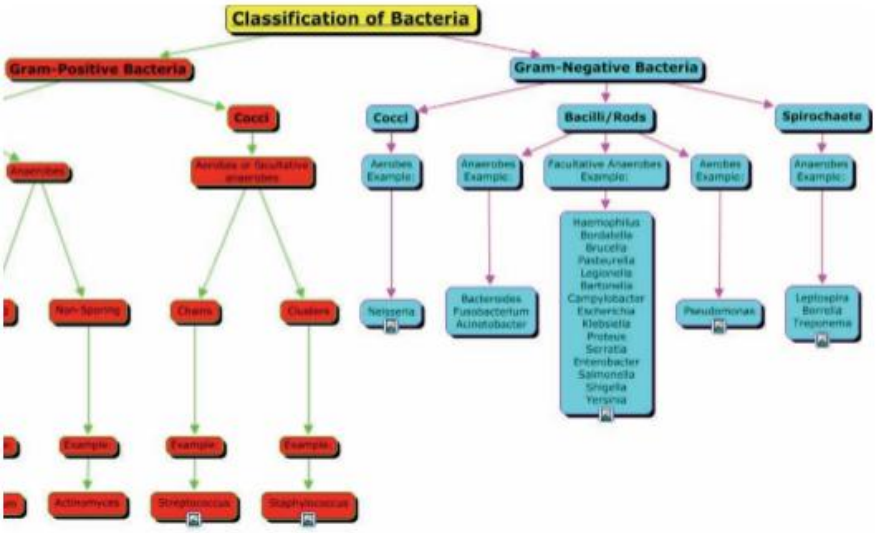


Figure I.5 : Exemple d'une taxonomie de bactéries

I.8.2. Thesaurus

Un thesaurus constitue un dictionnaire (vocabulaire contrôlé) hiérarchisé. Ce vocabulaire est normalisé et présente les termes (génériques ou spécifiques à un domaine) composant le vocabulaire sous une forme standard (lemme). Les termes y sont organisés dans une hiérarchie de concepts liés par des relations. Un thesaurus peut également fournir des définitions. Les relations couramment présentes dans un thesaurus sont des relations taxonomiques (hiérarchie), d'équivalence (synonymie), d'association (proximité sémantique, proche-de, relié-à, …). Il existe des thesaurus spécialisés dans des domaines précis tels que MeSH (domaine biomédical) Et WordNet qui est un thesaurus à caractère généraliste pour la langue anglaise dont l'organisation dépend du bon sens humain. Les noms, verbes, adjectifs et adverbes.

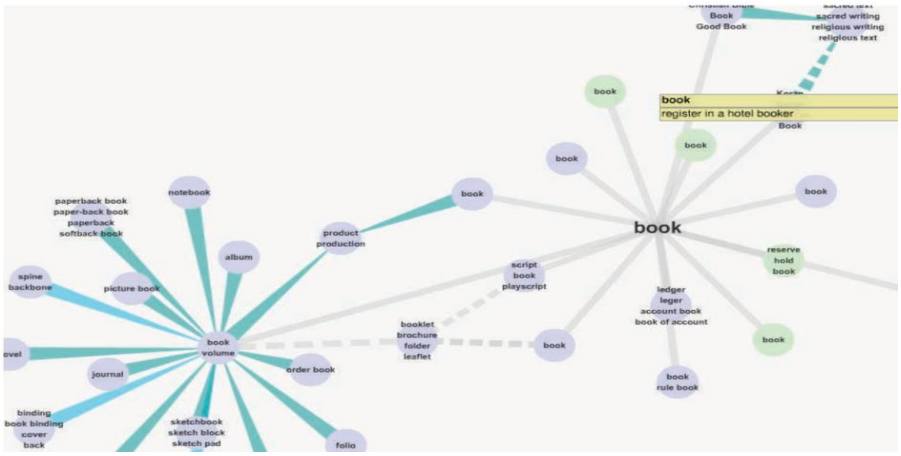


Figure I.6 : Le mot « book » dans le thésaurus anglais WordNet

I.8.3. Ontologie

Depuis les années 90, les ontologies sont devenues un des champs de recherche les plus populaires en informatique, investi par différentes communautés dont celle de l'intelligence artificielle (IA) et de la RI.

Les ontologies sont des spécifications conceptuelles d'un domaine. Elle définit et organise l'ensemble des concepts du domaine et les relations sémantiques entre eux. De cette façon, il est possible de donner à la machine un meilleur niveau de « compréhension » et il devient ainsi possible d'inférer des informations ce qui permet également de maintenir une plus grande cohérence.
Une des définitions de l'ontologie qui fait autorité est celle de Gruber : « Une ontologie est la spécification d'une conceptualisation d'un domaine de connaissance. » [Gruber 93].

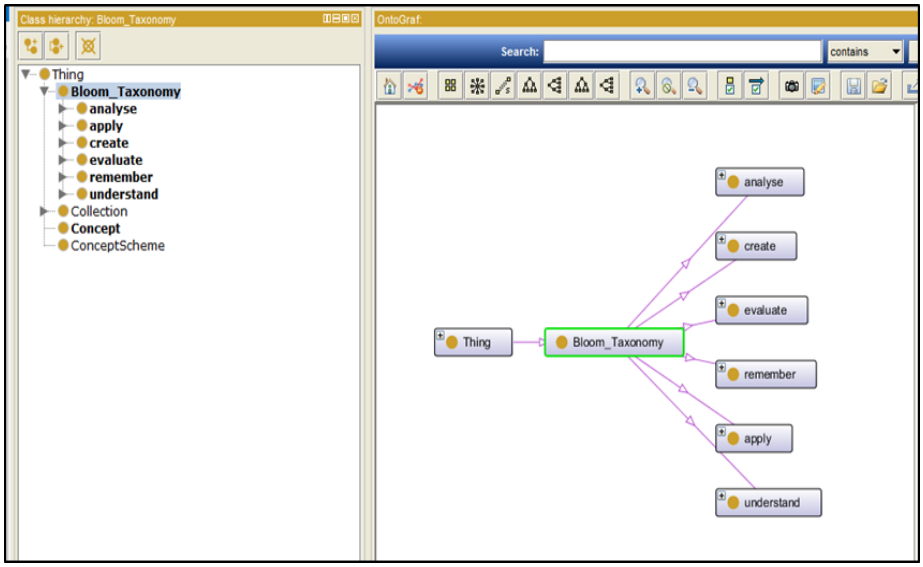


Figure I.7 : Exemple d'une ontologie

I .9. CONCLUSION

Nous avons présenté dans ce chapitre les principales notions et concepts de la recherche d'information. Nous y avons développé les principales étapes d'un processus de recherche d'information, à savoir, la représentation ou indexation de l'information, la comparaison de l'information et du besoin en information. Les principaux modèles existants dans la littérature ont été également présentés, ainsi que les différentes méthodes et cadres connus d'évaluation des performances des systèmes de recherche d'information. Ensuite, on a présenté les difficultés de la RI classique et que l'indexation sémantique est apparait pour pallier ces problèmes.

Dans le chapitre suivant, nous dressons un état de l'art sur l'utilisation des concepts de la sémantique en recherche d'information, et plus particulièrement dans les microblogs.

CHAPITRE II

La RI Sémantique

Dans

Les Microblogs

II.1. INTRODUCTION

Un service de microblogage est à la fois un moyen de communication et un système de collaboration qui permet le partage et la diffusion des messages textuels. Il permet aux utilisateurs de communiquer des informations sur leurs statuts, activités, pensées et opinions Au-delà de ces aspects d'utilisation personnelle à des fins de divertissement, les microblogs offrent aux entreprises et aux communautés virtuelles un moyen de collaboration rapide et pratique. En outre, les microblogs, vus comme une nouvelle source d'information, commencent à concurrencer les médias de masse.

En comparaison aux services de microblogage disponibles sur le Web, Twitter reste le site le plus populaire sur le web, 200 millions de microblogs ou tweets sont publiés chaque jour. 1,6 milliards de requêtes sont également émises chaque jour. Ainsi, les utilisateurs trouvent une difficulté pour accéder aux dernières actualités, masquées par l'énorme quantité des données et le flux soutenu des publications.

La recherche d'information dans les microblogs est particulièrement limitée par la taille courte des articles qui augmente à son tour la difficulté de la recherche textuelle par mots-clés. Les travaux dans ce domaine s'orientent alors vers l'intégration de la sémantique dans la recherche d'information.

Dans ce deuxième chapitre nous réalisons une recherche d'information sémantique pour pallier les problèmes rencontrés par les modèles de RI classiques, qui affectent la qualité de leurs résultats sur des corpus de microblogs : cas de Twitter. Dans un premier temps, nous présentons quelques spécificités des microblogs et celles de Twitter dans l'objectif de familiarisation.

La suite de cet article est organisée de la façon suivante : la section 3 dresse un accès à l'information dans les microblogs. Enfin, dans la section 4, nous présentons un état de l'art de la recherche d'information sémantique dans de Twitter.

II.2. SPECIFICATION DES MICROBLOGS

Un blog est un site web, sous forme d'un journal, où une ou plusieurs personnes appelées blogueurs publient leurs opinions et leurs analyses sur un événement actuel ou d'autres questions. Le texte publié peut contenir des liens hypertextes et plusieurs types de multimédias (images, vidéos, audio). Les blogs fournissent une forme d'interaction en ligne où les visiteurs peuvent lire le contenu du blog, laisser leurs commentaires, laisser des liens vers des informations supplémentaires, participer à des discussions avec les blogueurs et avec d'autres visiteurs. La discussion dans les blogs se fait d'une manière asynchrone (comme les courriels). Plusieurs plateformes de création des blogger[1], Skyrock Blog[2]) permettent une

[1] www.blogger.com/
[2] http://www.skyrock.com/blog/

diffusion simple et facile pour les blogueurs et les invités. Plusieurs travaux intéressés par l'analyse d'opinions ont eu recours aux blogs pour collecter les données.

Un article doit être pertinent et bien rédigé pour attirer l'attention des visiteurs. Cela implique qu'un article demande du temps pour le préparer. Pour cette raison, la plupart des internautes aiment mieux être passifs, ils préfèrent lire les articles des autres ou laisser des commentaires, que de créer des blogs et diffuser leurs propres articles. Généralement, les informations publiées dans les blogs ne sont pas en temps réel.

Une nouvelle tendance à émergé est celle du microblogage. C'est un dérivé du blog traditionnel qui permet aux utilisateurs de publier des messages courts (entre 140 et 200 caractères au maximum) sans titre, qui peut contenir également plusieurs types multimédias.

Au début, l'idée était de permettre aux internautes d'indiquer aux autres ce qu'ils sont en train de faire. Cependant, les choses se sont développées vite et les internautes ont profité de ce service pour exprimer leurs opinions sur différents sujets, diffuser des informations et faire des discussions. Contrairement aux blogs la plupart des internautes sont actifs, car ils n'ont plus besoin de rédiger de longs textes.

Le style d'écriture, employé dans les microblogs, est parfois incompréhensible par les non-initiés ou par les gens qui ne font pas partie de la conversation. Les utilisateurs commettent fréquemment des fautes d'orthographe et de grammaire, utilisent des abréviations, étirent des mots, et utilisent d'onomatopées (rire=haha) et des néographies (qui =ki). Plusieurs plateformes offrent le service de microbloggage tels que : Twitter.

Au départ, le service de microblogage était utilisé notamment par les jeunes, mais présentement tous les groupes d'âge utilisent ce service. Les microblogs (pour le microbloagage) sont devenus d'excellents outils pour des entreprises pour faire des publicités sur leurs produits et services ou pour les célébrités pour communiquer avec leurs fans. Les microblogs jouent aussi le rôle d'un réseau social, parce que les utilisateurs sont en mesure de faire des relations avec d'autres. On peut trouver deux types de relation dans les microblogs :

• Asymétriques : Un utilisateur A suit un utilisateur B sans que B suive A, cela implique que A peut consulter (sur son tableau de bord) les messages de B. Par contre, B ne peut pas consulter ceux de A.

• Symétrique: Un utilisateur A suit un utilisateur B et B suit A, cela implique que chacun peut consulter (sur son tableau de bord) les messages de l'autre.

Plusieurs raisons ont encouragé les internautes à s'inscrire à des microblogs : facilité d'utilisation, contact avec les amis, information en temps réel ou échange d'idées avec les autres. Les études les plus récentes, portant sur l'analyse des opinions et des sentiments, ont choisi les microblogs comme source des données vues le nombre important des messages publiés par jour. Les messages peuvent contenir beaucoup de sentiments et d'émotions parce que la majorité des messages sont publiés d'une façon spontanée.

II.3. PRESENTATION GÈNÈRALE DE TWITTER

Twitter	
Création	21 mars 2006
Personnages clés	Jack Dorsey, Noah Glass, Evan Williams, Biz Stone,
Action	NYSE : TWTR
Slogan	Quoi de neuf ? – Découvrez ce qui se passe en ce moment chez les personnes et dans les organismes qui vous tiennent à cœur – Suivez vos passions
Siège social	San Francisco, Californie (États-Unis)
Direction	Jack Dorsey (président), Dick Costolo (CEO, démissionne en juin 2015)
Activité	Internet
Produits	Service de microblog
Effectif	3 900 (2015)[1]
Site web	twitter.com

« Twitter est un outil de microblogage géré par l'entreprise Twitter Inc. Il permet à un utilistaeurs d'envoyer gratuitement de brefs messages, appélés *Tweets*, sur internet, par messagerie instantanée ou par SMS. » -Wikipédia-.

Figure II.1 : Interface d'utilisation Twitter

Le principe de Twitter est basé sur la seule phrase : « Quoi de neuf ? »

Les usagers de Twitter, des individus, entreprises, gestionnaires, consultants, organismes ou autres, répondent à cette question en utilisant un maximum de 140 caractères, et la réponse peut contenir un lien.

Le but est de communiquer avec les abonnés qui « suivent » votre profil.

Twitter compte environ 645 millions d'utilisateurs inscrits, dont plus de 300 millions actifs [3] qui publient plus de 60 millions de tweets quotidiennement.
Le but, sur Twitter, est de parvenir à rejoindre ce qu'on appelle les influenceurs [4] qui vont retweeter ou partager notre contenu à leurs nombreux abonnés.

Le « Hashtag » ou « mot-clic » est une invention propre à Twitter. Il s'agit d'un mot clé juxtaposé au symbole dièse (#) qui permet, en un seul clic, de trouver tous les tweets qui sont associés à ce sujet. L'utilisation des « Hashtag » dans les tweets permet aux usagers de trouver des publications qui les intéressent même s'ils ne sont pas abonnés au compte de la personne qui les a publiés. **Exemples de Hashtag** : #Algerie #Kabylie #Tiziouzou

Tous les « tweets » sont publics (à l'exception des comptes protégés) et sont la propriété de Twitter ; ils peuvent être trouvés au moyen du moteur de recherche Google.

Recherche des publications sur Twitter

Puisque tous les « tweets » sont publics, la fonction http://search.twitter.com permet de suivre tout ce qui se dit à votre sujet, ou au sujet de votre entreprise/organisme. Ceci permet de réagir rapidement en cas de crise (exemple : émission diffusée sur TF1 cet été).

Une application comme Google Reader[5] permet d'effectuer ce suivi quotidiennement et de répondre aux gens qui se posent des questions à notre sujet.

II.4. ACCES A L'INFORMATION DANS LES MICROBLOGS

L'accès à l'information consiste à analyser (souvent superficiellement) des textes pour en obtenir des informations en vue d'une application précise. Les travaux sont portés sur différents aspects, notamment: la classification thématique des microblogs, la recherche des microbloggeurs, la Recherche temps-réel de microblogs ainsi que l'analyse d'évènements, d'opinions et de tendances.

[3] Utilisateur actif à est celui qui suit 30 comptes et est suivi par un tiers d'entre eux.
[4] Influenceurs : personnalités connus, des journalistes, des blogueurs, des spécialistes des médias sociaux, etc.
[5] Google Reader : est un lecteur de flux d'informations au format RSS et Atom sur Internet.

II.4.1. Classification thématique des microblogs

L'objectif de la classification thématique de microblogs est de créer des filtres thématiques sur les flux d'information. Ceci est réalise en identifiant les sujets discutes dans les microblogs. La classification thématique des microblogs nous permettra, par extension, de classer les utilisateurs en fonction de leurs centres d'intérêts. [Damak 14].

Une première solution pour ce type de problème est de regrouper les microblogs en fonction des hashtags qu'ils contiennent [Efron 10]. Par ailleurs, les hashtags peuvent être extraits des résultats de recherche pour étendre ensuite la requête initiale.

Ramage et al. (2010) ont utilisé une implémentation étiquetée de LDA[6] (Latent Dirichlet Allocation) afin d'extraire des tags et de les utiliser pour caractériser les utilisateurs et les microblogs. Par ailleurs, des informations spatiotemporelles des tendances ont été exploitées dans [Song et al., 10] afin d'identifier les tags co-occurents. Ces tags sont utilisés par la suite pour regrouper les tweets dans des classes. Cependant, cette approche génère une distribution de termes liés thématiquement plutôt que des sujet significatifs.

Enfin, Bernstein et al. (2010) ont considéré quels sujets présentés sous forme de distributions de termes ne sont pas très utiles, et ont proposé un algorithme pour détecter précisément les sujets des microblogs. Ce dernier consiste à détecter les entités nommées dans un microblog et les soumettre à un moteur de recherche. Le sujet du microblog correspondra alors au terme le plus important dans les résultats, calculé à travers un algorithme de pondération (TF·IDF [Sparck Jones, 1988]).

II.4.2. Recherche de microbloggeurs

La recherche de microbloggeurs s'apparente a la tâche de recherche d'experts de la RI classique. Les objectifs sont l'identification des utilisateurs les plus populaires, ceux qui ont les mêmes centres d'intérêts que l'utilisateur courant, ou bien les experts dans des domaines spécifiques. [Damak 14].

TwitterRank [Weng et al., 10] est une approche inspirée de l'algorithme PageRank (voir annexe) [Brin et al., 1998] qui mesure la popularité des utilisateurs de Twitter.

II.4.3. Recherche temps-réel de microblogs

La recherche temps réel offre également aux internautes la possibilité d'interagir sur des documents de toute nature mis en ligne comme des textes, des photos, des cartes…. Pour

[6]LDA : est une technique statistique qui sert à détecter les sujets abstraits à partir d'une collection de documents

cette tache, l'utilisateur souhaite obtenir de l'information pertinente la plus fraiche possible vis-à-vis d'un besoin en information [Ounis et al., 11]. Généralement, un certain temps s'écoule avant que cette information soit disponible sur le web et qu'elle soit indexée par les moteurs de recherche [Dong, Zhang, et al., 10].

Dans la RI temps-réel, la date de publication d'un document est considérée comme un facteur de pertinence très important, si ce n'est pas le plus pertinent. Une interprétation possible de cette tache consiste a trier anti-chronologiquement tous les documents publies avant la date de soumission de la requête, et ensuite, à écarter les documents non pertinents [Ounis et al., 11] La tache se réduit donc à l'identification des caractéristiques des documents pertinents à restituer.

Plusieurs travaux ont proposé des critères utilisés comme facteurs de pertinence supplémentaires à la pertinence textuelle : la fraicheur [Magnani et al., 12 ; Vosecky et al., 12], la popularité de l'auteur [Zhao et al., 11 ; Massoudi et al., 11), la présence d'URL [Vosecky et al., 12]. . . Des études empiriques ont montre que ces critères reflètent la pertinence lorsqu'ils sont employés en plus de la pertinence textuelle [Damak et al., 13].

II.4.4. Détection d'événements

Twitter constitue un excellent moyen pour diffuser des informations, pour discuter des évènements et pour donner des avis. Plusieurs recherches ont montré que le contenu de ces outils reflète étroitement l'intérêt et les préoccupations des utilisateurs en temps réel. Un évènement est représenté par un ensemble de termes.

La détection d'événement a pour objectif de regrouper automatiquement les termes représentant un même sujet, et de trouver les sujets (évènements) les plus importants (fréquence des termes, etc.).

II.4.5. Détection d'opinions

La détection d'opinion ou l'analyse de sentiments a été souvent étudiée en recherche d'information, particulièrement dans la recherche de blogs [Pang et Lee, 08 ; Missen et al., 09]. L'objectif de la détection d'opinions est de retrouver les documents exprimant des opinions sur le sujet de la requête.

Plusieurs travaux [Yu et Hatzivassiloglou 03], [Wiebe et al. 01] et [Dave et al. 03] ont montré qu'il faut s'assurer que le texte exprime une opinion avant de déterminer son type global. On parle souvent de polarité de texte dans ces cas : polarité positive correspondant à une opinion favorable et négative à défavorable. Pour certaines études, on utilise d'autres classes telles qu'excellent, très bien, bien, moyennes ou neutres.

II.4.6. Détection de tendances

La détection de tendances vise à identifier automatiquement les thèmes émergeant qui apparaissent dans le flux de microblogs en temps-réel [R. Li et al., 12].

Les tendances sont généralement des événements émergents, les dernières nouvelles et les sujets qui attirent l'attention des utilisateurs. La détection des tendances revêt donc une grande utilité pour les journalistes et les analystes, car elle leur permet d'être rapidement actifs sur les sujets « tendances ».

II.5. ETAT DE L'ART

Twitter constitue une source continue et illimitée de données en langage naturel qui est particulièrement difficile à traiter avec les approches classiques de traitement automatique du langage naturel (TAL) car le type de langage utilisé dans Twitter est très éloigné des normes du langage traditionnel, avec ses conventions (hashtags, mentions, retweet...), son lexique particulier (souvent grossier, avec des abréviations/ émoticons/ acronymes omniprésents...) et sa syntaxe parcellaire dans le meilleur des cas. Pourtant, Twitter est l'un des mediums de communication contemporain les plus utilisés et les plus riches en contenu sémantique.

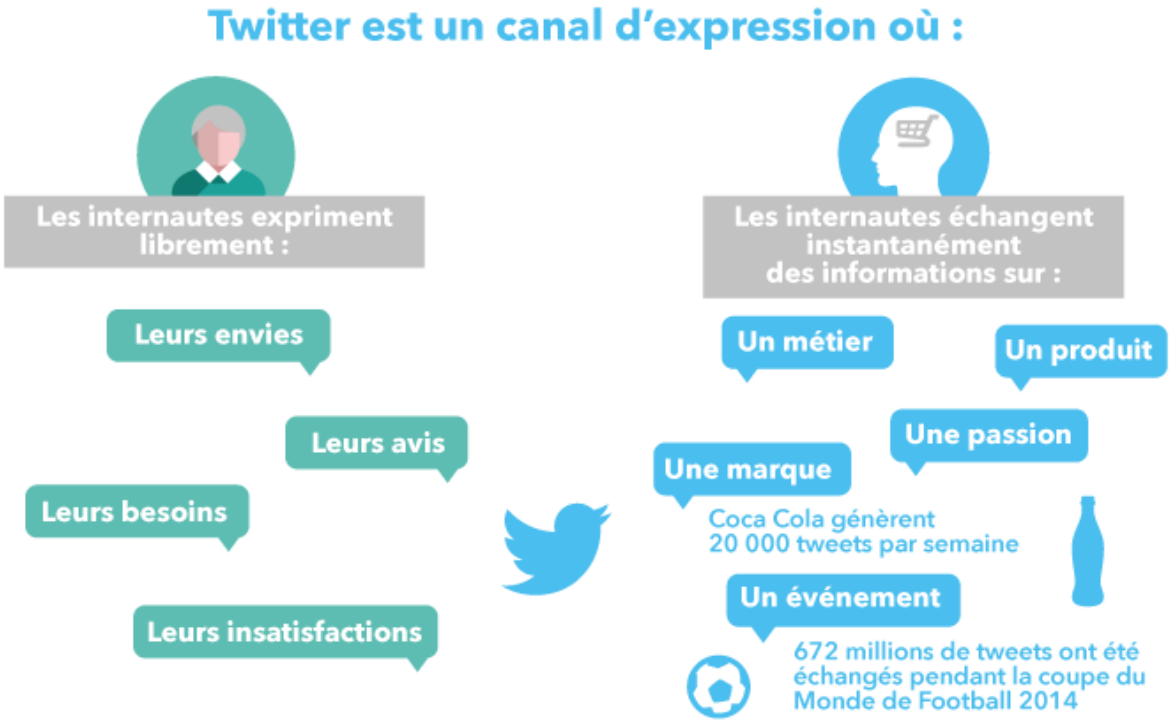


Figure II.2 : Twitter, un canal d'expression.

Travaux de RI sémantique centré sur le contenus des messages microblogs

Généralement, toutes les approches qui tentent d'extraire de l'information des microblogs cherchent à introduire la sémantique pour compenser l'impuissance des systèmes de recherche classique. Ainsi, plusieurs chercheurs ont proposé d'exploiter le contenue des messages microblogs 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 centrées sur le contenu des messages microblogs avec utilisant des méthodes différentes.

Approche basée sur l'enrichissement sémantique

Cette approche consiste à ajouter la sémantique à des tweets en identifiant automatiquement les concepts qu'ils leurs sont sémantiquement liés et à générer des liens vers les articles de Wikipedia correspondant en outre, on prend un concept comme un élément qui a une entrée unique et sans ambiguïté dans Wikipedia.

Le but de cette méthode est d'obtenir un haut score rappel/précision. On augmente le rappel en générerons une liste des concepts candidats pour chaque mot dans un tweet en utilisant les approches comme l'appariement lexical. Et on améliore la précision en déterminant le quel des concepts candidats à garder par l'application de l'apprentissage supervisé. (Edgar, Weerkamp et Rijke [1]).

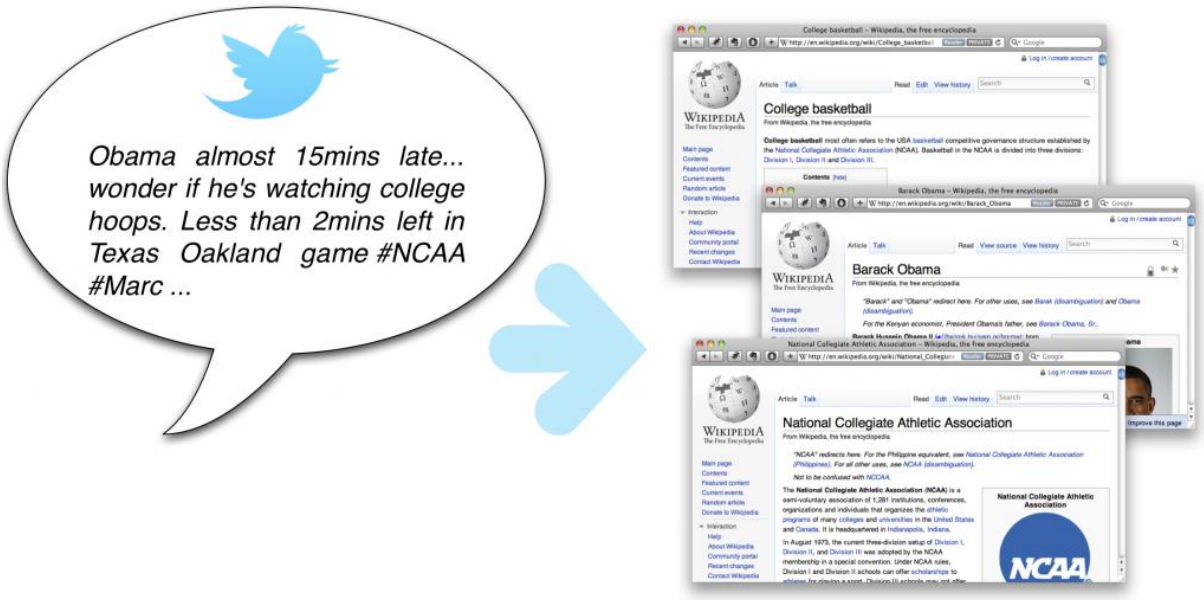


Figure II.3: illustration de l'approche d'Edgar Meij, Wouter Weerkamp, Maarten de Rijke.

Approche basée sur les relations sémantique

Cette approche propose une nouvelle technique « *the Wikipedia Link-based Measure*», un algorithme qui calcule la proximité sémantique entre les termes en utilisant les liens trouvés dans leurs articles Wikipedia correspondant. Contrairement à d'autres techniques basées sur Wikipedia, WLM est en mesure de fournir des mesures précises de manière efficace, en utilisant uniquement les liens entre les articles plutôt que leur contenu textuel (Milne et Witten [2]).

Approche basée sur l'annotation sémantique

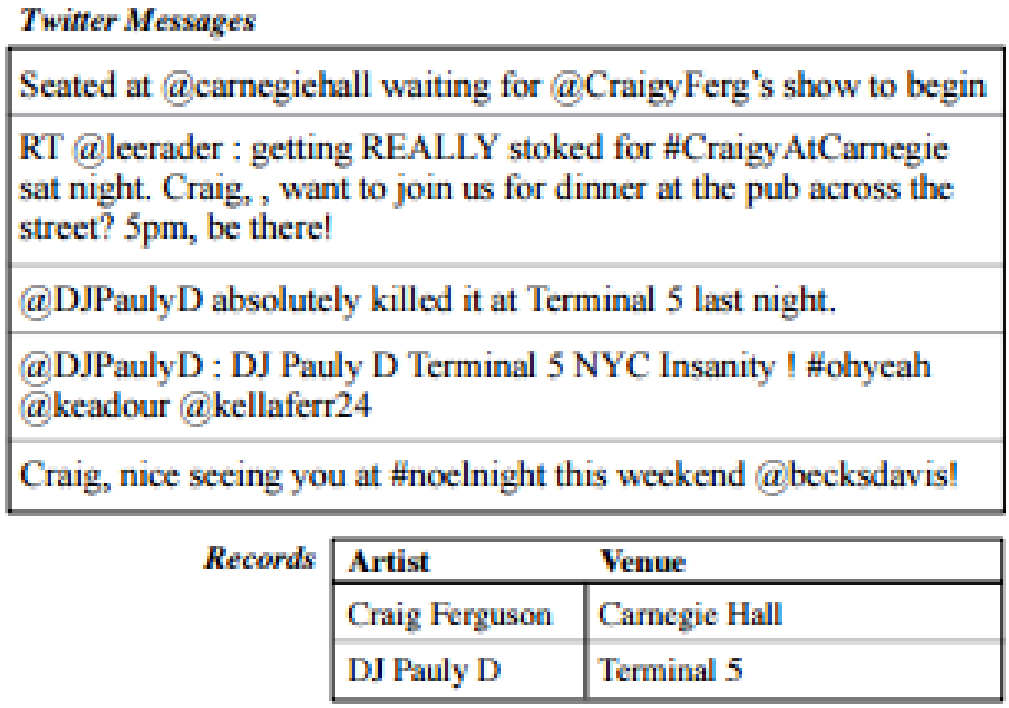
Dans cette approche on met en œuvre un TAGME, un système qui est en mesure d'augmenter efficacement et judicieusement un texte clair avec des hyperliens pertinents aux pages de Wikipédia. La spécialité de TAGME par rapport aux systèmes connus [4, 5] est qu'il peut annoter les textes qui sont courtes et mal composé, tels que des extraits de résultats des moteurs de recherche, tweets, presses, etc .. Cette annotation est extrêmement instructive, donc une tâche qui est actuellement traité en utilisant le modèle sac-de-mots (voir annexe) pourrait bénéficier de l'aide de cette annotation et tirer parti des millions de pages Wikipédia et de leurs inter-relations (Ferragina et Scaiella [3]).

Approche se basant sur la suggestion de requêtes

La transformation de requête est une stratégie souvent utilisée dans les moteurs de recherche pour obtenir des requêtes qui sont en mesure de retourner des résultats de recherche plus utiles que la requête d'origine et les moteurs de recherche les plus populaires fournissent des installations qui permettent aux utilisateurs de compléter, de préciser ou de reformuler leurs requêtes. Cette approche étudie le problème de la suggestion de requête, un type spécial de transformation, (Meij et Bron [6]).

Approche basée sur les enregistrements

Cette approche propose une nouvelle méthode pour l'extraction d'enregistrement à partir des flux sociaux tels que Twitter. Contrairement à des configurations d'extraction typiques, ces environnements sont caractérisés par des messages courts avec un discours très familier. E. Benson, A. Haghighi, et R. Barzilay développent un modèle graphique qui aborde ces problèmes en apprenant un ensemble d'enregistrements et un enregistrement de message aligné simultanément; la sortie de leur modèle est un ensemble d'enregistrements canoniques, dont les valeurs sont compatibles avec les messages alignés, (Benson et al. [7]).



Twitter Messages

Seated at @carnegiehall waiting for @CraigyFerg's show to begin
RT @leerader : getting REALLY stoked for #CraigyAtCarnegie sat night. Craig, , want to join us for dinner at the pub across the street? 5pm, be there!
@DJPaulyD absolutely killed it at Terminal 5 last night.
@DJPaulyD : DJ Pauly D Terminal 5 NYC Insanity ! #ohyeah @keadour @kellaferr24
Craig, nice seeing you at #noelnight this weekend @becksdavis!

Records

Artist	Venue
Craig Ferguson	Carnegie Hall
DJ Pauly D	Terminal 5

Figure II.4: illustration de l'approche d'E. Benson, A. Haghighi, et R. Barzilay

II.6. CONCLUSION

Dans ce chapitre, nous avons présenté les principales notions qui s'intéressent à la recherche d'information sociale, les spécificités des microblogs et de la plateforme Twitter ainsi que les principaux moyens d'accès à l'information dans les microblogs, pour finir nous avons exploré un état de l'art de la recherche d'information sémantique dans Twitter. Dans la partie suivante nous détaillerons l'approche que nous avons proposée pour enrichir les différentes approches de recherche d'information classique proposées pour pallier aux différents problèmes de cette dernière.

CHAPITRE III

Approche Proposée

III.1. INTRODUCTION

Dans le chapitre précédent sur l'état de l'art de la recherche d'information sémantique dans les microblogs, Nous avons passé en revue de nombreuses approches existantes. Ces approches proposent d'exploiter plusieurs critères afin de résoudre cette problématique dans TWITTER, certaines se basent sur l'étude de TWITTER comme un réseau social et analyse les interactions entre les individus et d'identifier les acteurs influents et susceptibles de produire des informations pertinentes, ou sur la suggestions de requêtes [6] afin d'aboutir à des résultats plus précis, tandis que d'autres approches se basent sur le contenue et proposent d'enrichir le contenu des tweets de sorte à ce que les tweets les plus pertinents soient retournés en premier [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 inspirons de l'approche d'Edgar [1].

III.2. DESCRIPTION DE L'APROCHE PROPOSÈE

Nous proposons une approche inspiré de celle d'Edgar (Voir Etat de l'art chapitre2) , à savoir déterminer le sujet d'un message microblog en identifiant automatiquement des concepts dans les tweets que nous prenons pour être n'importe quel article qui a une entrée unique et sans ambiguïté dans une source de connaissance célèbre à grande échelle Wikipedia (voir annexe).

La jonction du texte aux ressources de connaissance, d'autre part, a reçu une quantité croissante d'attention ces dernières années. En commençant du domaine de reconnaissance d'entité nommée (NER), Au lieu de simplement identifier des types, nous avons aussi pour but de désambiguïser les concepts trouvés et les lier avec des articles de Wikipedia. Avec plus de 3.5 millions d'articles, Wikipédia est devenu une source riche de connaissance et une cible pour la jonction; l'utilisation d'approches de jonction automatique utilisant Wikipedia a rencontré le succès considérable [2].

Dans la première partie de notre travail, Nous présentons une méthode d'annotation pour automatiquement détecté les concepts portés par chaque tweet et les relier à Des articles de Wikipedia.

Pour finir on extrait automatiquement de chaque article Wikipedia liée de nouveaux concepts qui serviront à l'enrichissement de nos tweets pour faciliter l'extraction d'informations de ces derniers sur un niveau sémantique.

Notre approche proposée implique une méthode à deux étapes pour la jonction sémantique.

- La première étape est orientée Rappel où le but est d'obtenir une liste classée des concepts candidats.

- Dans la deuxième étape, nous améliorons (augmentons) la précision et décidons lequel des concepts candidats à garder pour l'enrichissement.

Notre deuxième partie d'évaluation concerne une comparaison de deux méthodes (classique et sémantique) pour notre concept initial la recherche d'information dans les microblogs.

III.2.1. Désambigüisation

Les liens vers une structure de connaissances sont souvent considérés comme un moyen pour fournir une sémantique à des éléments numériques. Une approche simple et fréquemment prise pour relier le texte à des concepts est d'effectuer l'appariement lexicale entre parties du texte et les titres de concepts, Toutefois la simple correspondance entre un texte d'entré avec des titres de concepts souffre de plusieurs inconvénients y compris l'ambigüité, ces problèmes peuvent être résolus sur Wikipedia grâce à ses pages de désambigüisations et à DBpedia Spotlight qui nous fourni un moyen d'attribuer pour chaque concept identifié une entré unique reconnu par rapport au contexte du document dans cette source de connaissance.

III.2.2. Twitter

Divers auteurs ont tenté de « donner une signification » au texte en général [6] ou au texte contenu dans les tweets [7]. Notre approche est plus générale et vise à enrichir les tweets plutôt que l'extraction d'informations de ces derniers. Notre approche pourrait donc simplifier les tâches proposées.

III.2.3. Enrichissement sémantique

Étant donné le contenu des tweets, notre défi consiste à extraire des valeurs de la sémantique du contenu textuel. Plus loin, lors du traitement de notre approche, elle consiste à extraire le contenu principal du tweet. Pour cela, nous utilisons *dbpedia-spotlight* (voir annexe), une bibliothèque qui cherche des entités (concepts) connus dans le texte et tente de les relier à leurs identifiants uniques globaux dans *DBpedia*. Afin de soutenir la modélisation de l'utilisateur et la personnalisation, il est important de donné le contenu brut des tweets. Nous utilisons donc des prétraitements Qui permettent l'extraction précise d'entités et nous leurs attribuons des articles wikipedia unique. Les connexions entre les tweets et les articles wikipedia sont sémantiquement enrichies et nous permettent de construire un tweets enrichi dans un contexte sémantiquement bien défini.

III.2.4. Facteurs de pertinence

Nous partons de l'hypothèse qu'un tweet est pertinent en fonction de la méthode de « similarité cosinus », qui calcule la proximité sémantique (également nommée similarité sémantique) entre une requête et un Tweet.

Le rang d'un tweet dans le classement global des résultats sera donc déterminé par son score TF*IDF, qui est calculé en appliquons une relation entre un tweet, et un ensemble de tweets partageant des similarités en matière de mots clés. On recherche en quelque sorte une relation de quantité / qualité sémantique à travers un ensemble de tweets, notre but est de fournir des résultats aussi pertinent que possible. Ainsi Notre fonction de classement est définies comme suit :

$$ScoreTF*IDF\ (Q, Ti) = TF\ (Q, Ti) * IDF(Q)$$

- **Fréquence du terme**

La fréquence d'un terme (term frequency) est simplement le nombre d'occurrences de ce terme dans le document (Tweet) considéré.

$$TF\ (Q, Ti) = |n_{Q,\ Ti}\ | / |n_{Ti}\ |\ \ où:$$

• $|\ n_{Q,\ Ti}\ |$: nombre d'occurrence du terme Q dans le tweet Ti ;
• $|\ n_{Ti}\ |$: nombre de terme du Tweet Ti.

- **Fréquence inverse de document (tweet)**

La fréquence inverse de document (inverse document frequency) est une mesure de l'importance du terme dans l'ensemble de la collection de tweets.
Dans le schéma TF-IDF, on vise à donner un poids plus important aux termes les moins fréquents, considérés comme plus discriminants. Elle consiste à calculer le logarithme de l'inverse de la proportion du tweet de la collection qui contient le terme:

$$Idf(Q) = log\ |D| / |\ \{\ N_{Ti} : Q \in N_{Ti}\ \}\ |\ où:$$

• | D | : nombre total de Tweets dans la collection;
• | { N_{Ti} : Q / N_{Ti} } |: nombre de Tweet où le terme Q apparaît.

III.2.5. Mesures d'évaluation

Pour valider l'efficacité de notre approche À partir de l'ensemble de réponses obtenues, on peut mesurer les performances de notre approche mis en œuvre pour retrouver d'une manière précise tout les tweets pertinents en relation avec la requête. Les critères de mesure des performances qu'on a sélectionnées sont le rappel et la précision. Ainsi le score final est retourné par la fonction F-mesure qui une mesure populaire qui combine la précision et le rappel est leur moyenne harmonique, nommée F-score :

$$F\text{-}score = 2.\ (Précision * Rappel) / Précision + Rappel)$$

- **Précision**

La précision est le nombre de Tweets pertinents retrouvés rapporté au nombre de Tweets total proposé par la mise en œuvre de l'approche pour une requête donnée.

Le principe est le suivant : quand on test une requête, on souhaite que les tweets proposées en réponse à notre interrogation correspondent à notre attente. Tous les tweets retournés superflus ou non pertinents constituent du bruit. La précision s'oppose à ce bruit documentaire. Si elle est élevée, cela signifie que peu de tweets inutiles sont proposés par notre approche et que cette dernière peut être considérée comme "précise". On calcule la précision avec la formule suivante :

$$Précision = \frac{nbre\ de\ Tweets\ pertinents\ retrouvé}{nbre\ de\ Tweets\ total\ proposé\ par\ l'approche\ \ pour\ une\ requête\ donnée}$$

- **Rappel**

Le rappel est défini par le nombre de tweets pertinents retrouvés au regard du nombre de tweets pertinents que possède la collection. Cela signifie que lorsque on test une requête on souhaite voir apparaître tous les tweets qui pourraient répondre à son besoin d'information. Si cette adéquation entre la requête et le nombre de tweets présentés est importante alors le taux de rappel est élevé. À l'inverse si la collection possède de nombreux documents intéressants mais que ceux-ci n'apparaissent pas dans la liste des réponses, on parle de silence. Le silence s'oppose au rappel. On calcule le rappel avec la formule suivante :

$$Rappel = \frac{nbre\ de\ Tweets\ pertinents\ retrouvé}{nbre\ de\ Tweets\ pertinents\ à\ retrouver}$$

III.2.6. Architecture de l'approche proposée

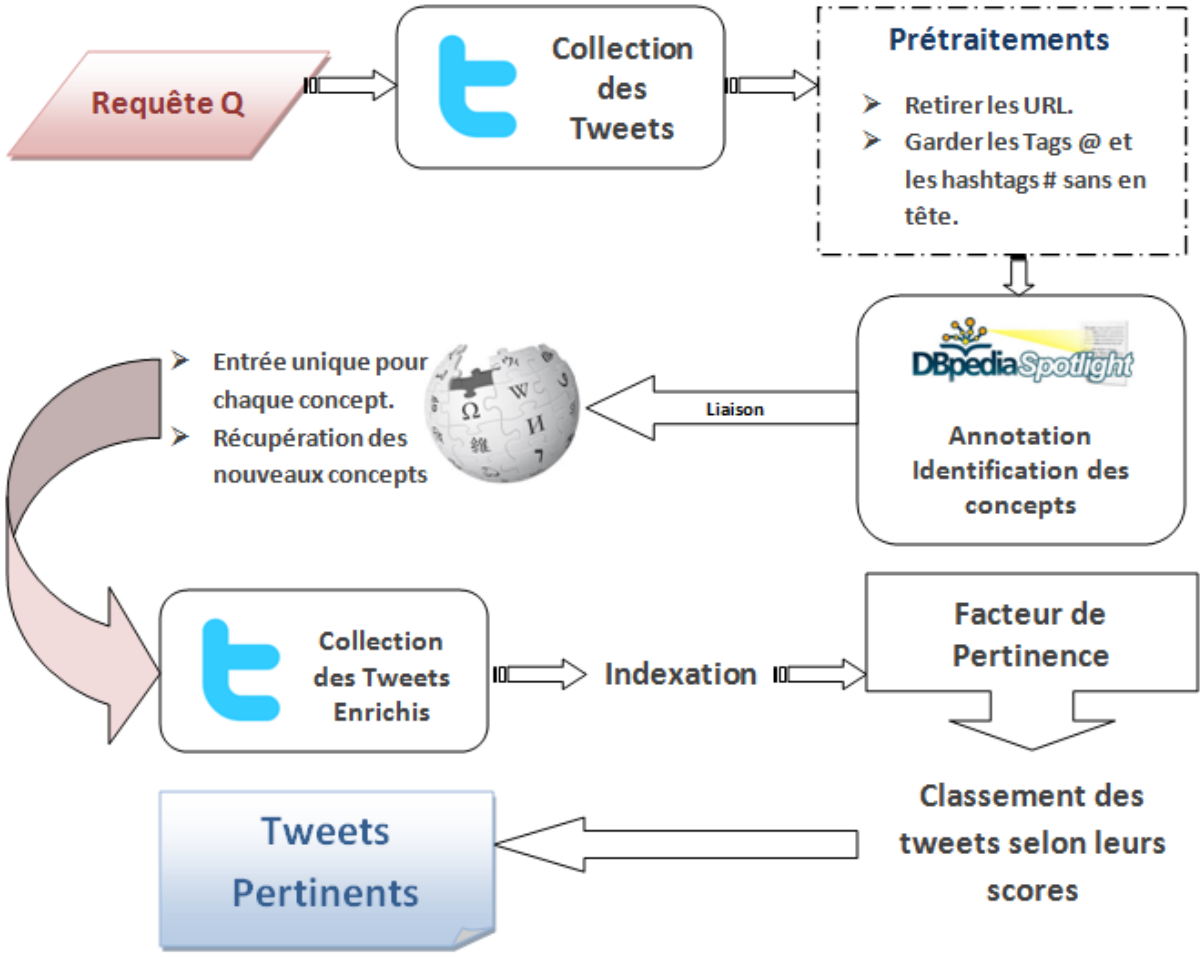


Figure III.1 : Ajout de la sémantique dans les messages microblogs (Tweet)

III.3. CONCLUSION

Dans ce chapitre nous avons décrit en détail notre approche, nous avons exploré en détail les travaux connexes utilisé par notre approche, nous avons décrit toutes les fonctions proposées dans notre méthode, on a clôturé le chapitre avec une figure qui représente au mieux notre approche dans le prochain chapitre, nous allons voir l'implémentation de notre ainsi que l'évaluation de ses résultats.

CHAPITRE IV

Implémentation Et Tests

IV.1. INTRODUCTION

Afin d'implémenter et de tester notre approche décrite dans le précédant chapitre, nous avons conçu plusieurs expériences. Dans ce dernier chapitre, nous détaillons notre cadre expérimental, y compris la façon dont nous prélevons les tweets, les annotations dont la version Wikipedia, DBpedia spootlight et la façon dont nous reliant les tweets à des concepts (Jonction) indispensable à l'enrichissement des ces tweets. Enfin nous conclurons ce chapitre par une comparaison des résultats obtenues par rapport à une autre méthode de recherche dans le but d'évaluer l'efficacité de notre approche.

IV.2. IMPLEMNTATION

Nous avons implémenté notre approche sous JAVA (voir annexe) en utilisant la bibliothèque LUCENE (voir annexe), nous avons développé ce projet sous l'ide ECLIPSE NEON (voir annexe). Dans notre implémentation nous avons annoté et liés les concepts à travers l'API DBpedia Spootlight.

IV.3. EXPERIMENTATION

IV.3.1 Protocole expérimental

Notre évaluation va se faire comme suit : nous allons effectuer une série de tests avec notre approche puis nous comparerons ceux-ci avec les résultats retournés par la méthode de recherche classique propre à LUCENE que nous avons implémentée. Pour rappel, notre fonction de classement est sous la forme :

$$ScoreTF*IDF\ (Q, Ti) = TF\ (Q, Ti) * IDF(Q)$$

LUCENE se base sur le même modèle de restitution, la fonction de classement est alors définie comme suit :

$$TF*IDF\ (Q, Ti) = TF\ (Q, Ti) * IDF(Q)$$

IV.3.2. Collection de Tweets

Sur un échantillon de 2000 tweets, Pear Analytics (voir annexe) classes comme 40% contenant "bavardage inutile," et avec 37,55% comme simple conversation. Ainsi, une partie importante de tous les tweets sont non informative et seule une petite fraction contient des sujets d'intérêt généraux. Afin de tenir compte de cette partialité possible, nous

échantillonnons au hasard 10 des utilisateurs des comptes Twitter vérifiés via l'API *TWITTER CURATOR* (voir annexe) (figure I.1) et pour chacun de ces utilisateurs, nous récupérons le derniers tweet. Les utilisateurs de Twitter dans cette liste peuvent être considérés comme influents et leurs tweets sont plus susceptibles d'être pris par d'autres utilisateurs de Twitter que d'un échantillon aléatoire d'utilisateurs de Twitter. Comme prétraitement, nous enregistrons toutes les URL, les mentions, et les hashtags dans chaque tweet. Les URL sont retirés de chaque tweet, tandis que les mentionnes (tag) et les hashtags sont gardés sans en tête, respectivement '@' et '#'.

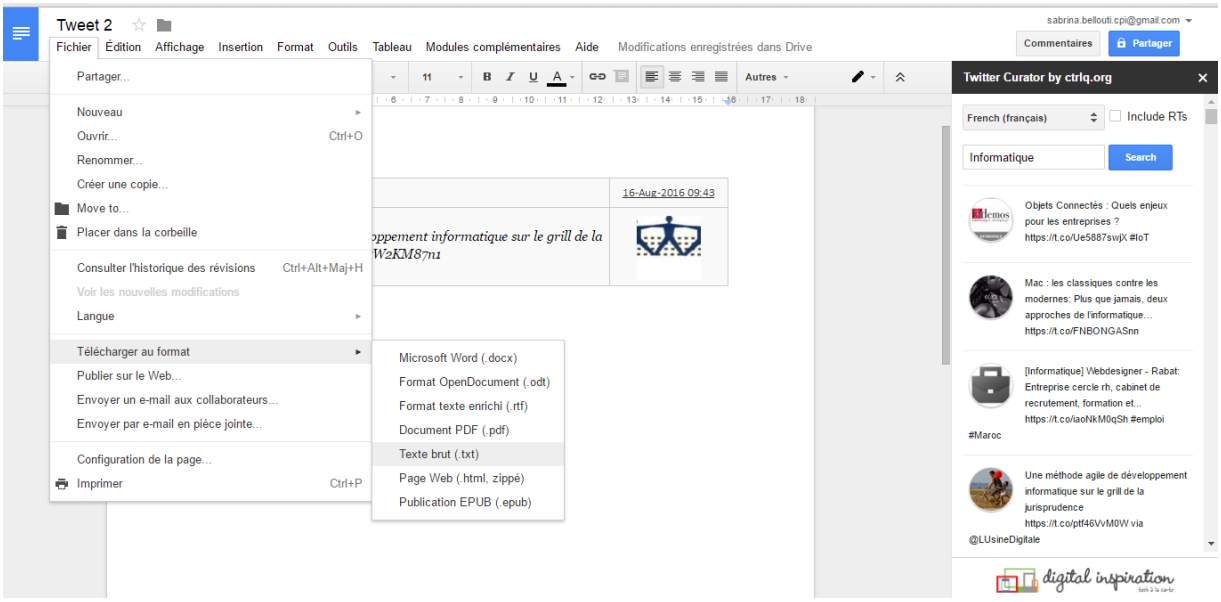


Figure IV.1 : Extraction de tweets via TWITTER CURATOR

IV.3.3. Wikipedia

Pour relier les tweets nous utilisons un dump[1] Wikipedia en ligne (courant) qui est daté du 11 Août 2016. Dans cette snapshot[2] particulière, nous avons 3,483,213 articles appropriés, 4,526,685 redirections, 120.547 pages de désambiguïsation (figure IV.2) et 71,204,142 hyperliens entre les articles. Pour les Ancres de lien[3], nous incluons non seulement les textes d'ancrage trouvés dans les liens interne [4]de Wikipedia, mais aussi les titres de toutes les pages de redirection pointant à un article.

[1] Dump : une sauvegarde de fichier et une sauvegarde de bases de données à une instante donnée de l'état de ses bases. Wikipédia propose des dumps de son contenu à l'adresse : http://download.wikimedia.org/
[2] Snapshot : version statique (donc non modifiable par les écritures ultérieures) de toute ou partie de ce que stocke une BDD, un système de fichier ou un disque.
[3] Ancre de lien : texte cliquable qui apparait dans un lien hypertexte
[4] Un lien interne : est un lien vers un autre article Wikipedia.

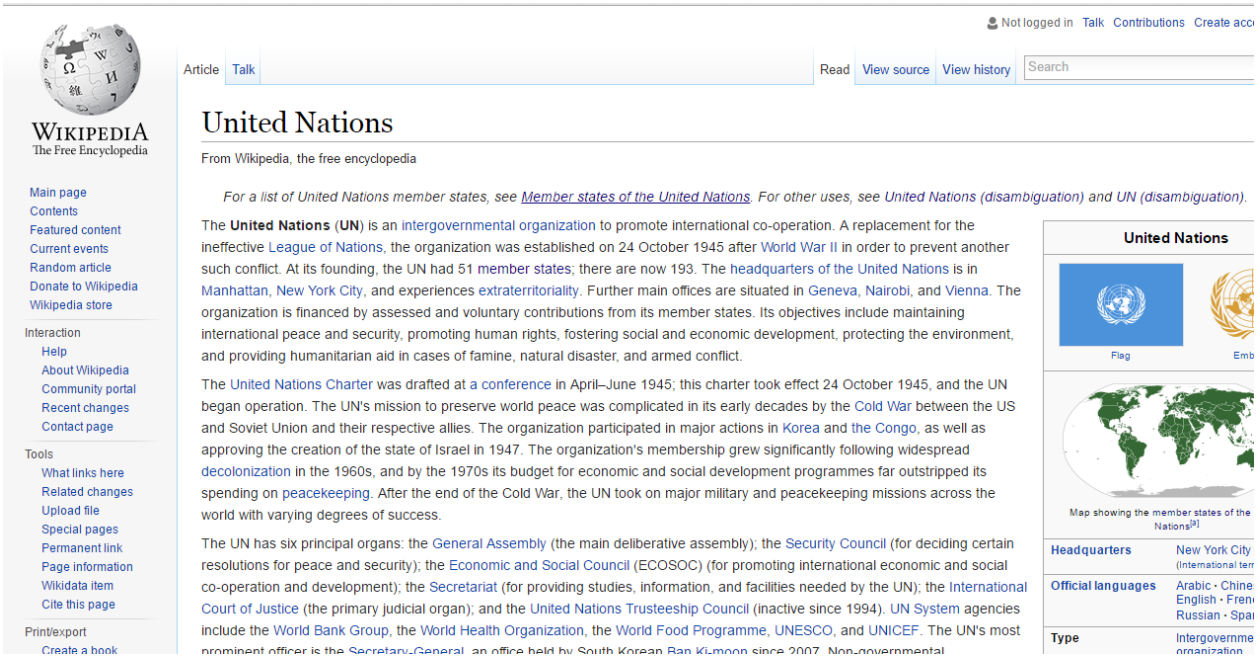


Figure IV.2 : Exemple d'une Page de désambigüisation Wikipedia

IV.3.4. Annotation

Afin d'obtenir des annotations, nous avons utilisé l'API DBpedia Spootlight pour annoter 10 tweets, chacun contenant 15 termes en moyenne. DBpedia spootlight est une interface d'annotation avec laquelle on va identifie les concepts contenus dans, signifiés par, ou pertinentes pour chaque tweet. Elle indique également si un tweet est soit ambiguë (où plusieurs concepts cibles existent) ou erronées (en l'absence de concept pertinent la sémantique pourrait être affectée). DBpedia Spootlight à identifiés 2 .5 concepts par tweet en moyenne.

Dans le fichier 'Tweets Annotation', vous trouverez une liste des concepts annotés pour chaque tweet. Ainsi chacun de ces concepts représente une entré unique et sans ambigüité vers un titre de l'article de Wikipedia.

Le fichier 'DataS' est un fichier contenant les tweets enrichis avec les nouveaux concepts cibles extrait du dump Wikipedia.

IV.4. RESULTATS ET DISCUSSION

Dans cette section, nous présentons les résultats obtenus lors de l'expérimentation de notre approche et de la recherche classique (considérée comme Baseline) que nous avons

implémentée, puis nous comparons ces résultats et nous les analyserons. Nous récapitulons les résultats dans des tableaux et des graphiques.

IV.4.1. Comparaison entre nos résultats et ceux de la recherche classique

Une fois un ensemble de documents potentiels identifiés comme pouvant répondre à une requête, il s'agit de les ordonner par ordre de pertinence.

	Tweet 1	Tweet 2	Tweet 3	Tweet 4	Tweet 5	Tweet 6	Tweet 7	Tweet 8	Tweet 9	Tweet 10	Score TF*IDF
Q 1	0	0	**0**	0	0	0	0	0	0	0	**0**
Q 2	0	0	0	0	0	0	0	0	**0**	0	**0**
Q 3	**0,170**	**0,213**	**0 ,128**	0	0	0	0	0	**0**	0	**0,512**
Q 4	0	0	0	0	0	0	0	0	0	**0**	**0**
Q 5	0	0	0	0	**0**	0	0	0	0	0	**0**
Q 6	0	**0,199**	0	0	0	0	0	0	0	0	**0,199**

Tableau IV.1 : Les valeurs TF*IDF obtenues dans la recherche classique

	Tweet 1	Tweet 2	Tweet 3	Tweet 4	Tweet 5	Tweet 6	Tweet 7	Tweet 8	Tweet 9	Tweet 10	Score TF*IDF
Q 1	0	0	**0,119**	0	0	0	0	0	0	0	**0,119**
Q 2	0	0	0	0	0	0	0	0	**0,217**	0	**0 ,217**
Q 3	**0 ,879**	**0,659**	**0,395**	0	0	0	0	0	**0,479**	0	**2 ,415**
Q 4	0	0	0	0	0	0	0	0	0	**0,184**	**0,184**
Q 5	0	0	0	0	**0,299**	0	0	0	0	0	**0,299**
Q 6	0	**0,399**	0	0	0	0	0	0	0	0	**0,399**

Tableau IV.2 : Les valeurs TF*IDF obtenues dans la recherche Sémantique

	Q3 = « Computing »	
Tweets	**Notre Approche**	**Approche Classique**
Tweet1	**0 ,879◄**	**0,170**
Tweet2	**0,659**	**0,213◄**
Tweet3	**0,395**	**0 ,128**
Tweet9	**0,479**	**0**

Tableau IV.3 : Comparaison du classement par pertinence des deux approches

Nous avons déterminé après des séries de tests que le Tweet1 était le Tweet le plus pertinent pour la requête **Q3**. Nous remarquons dans le tableau ci-dessus (Tableau IV.3) que

sur la requête **Q3**, la fonction de classement de notre approche classe le Tweet1 comme étant le plus pertinent où la recherche classique classe le Tweet2. Conformément à notre classement initial notre approche retourne un bon résultat en termes de pertinence par rapport à la Baseline.

IV.4.2. Evaluation

Une fois un nombre de documents potentiels retrouvé par le système comme pouvant répondre à une requête, il s'agit dés lors d'évaluer le système en terme de précisons et de rappel.

	Précision	
Intitulé de la requête	**Notre Approche**	**Approche Classique**
Administrative court	**1**	**0**
Cybersecurity	**1**	**0**
Computing	1	1
Automobile	**1**	**0**
Agile software development	**1**	**0**
History of computing	**1 (99%) ▲**	**0,33 (3%) ▼**

Tableau IV.4 : Précision des deux approches

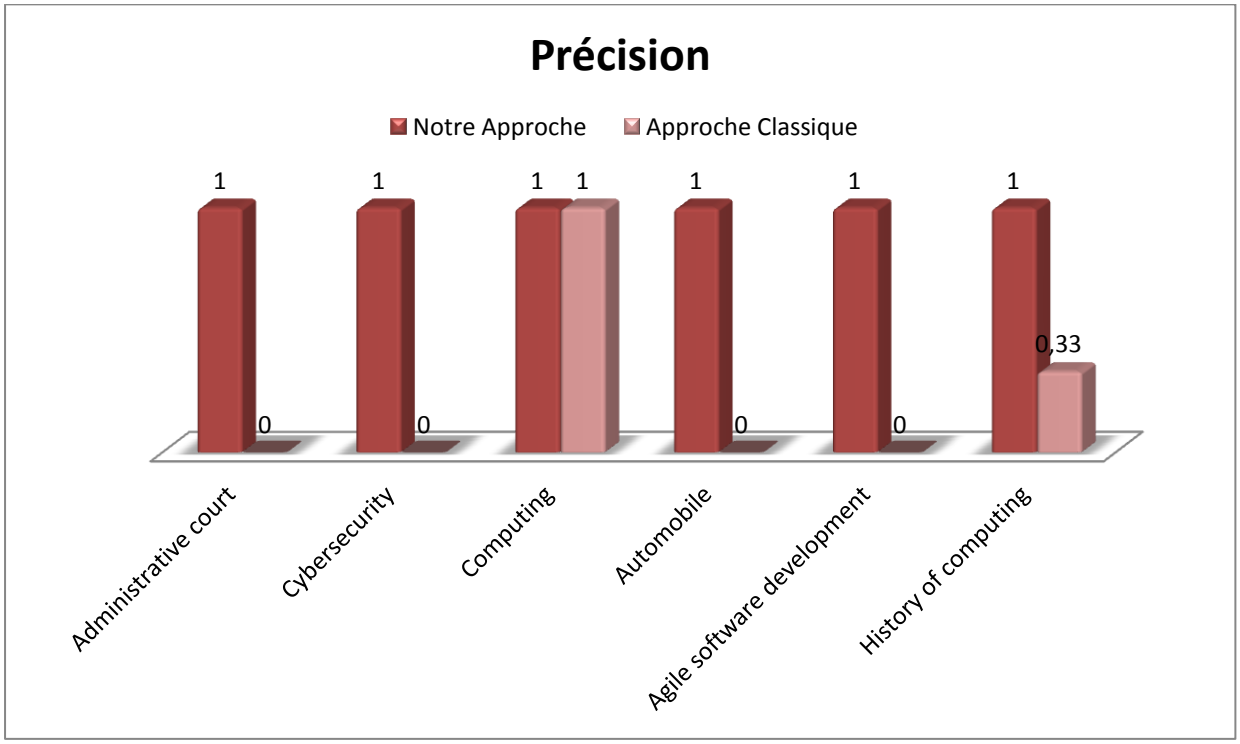


Figure IV.3 : Comparaison de la précision entre les deux approches

	Rappel	
Intitulé de la requête	**Notre Approche**	**Approche Classique**
Administrative court	**1**	**0**
Cybersecurity	**1**	**0**
Computing	**0,8 (80%) ▲**	**0,6 (60%) ▼**
Automobile	**1**	**0**
Agile software development	**1**	**0**
History of computing	**1**	**1**

Tableau IV.5 : Rappel des deux approches

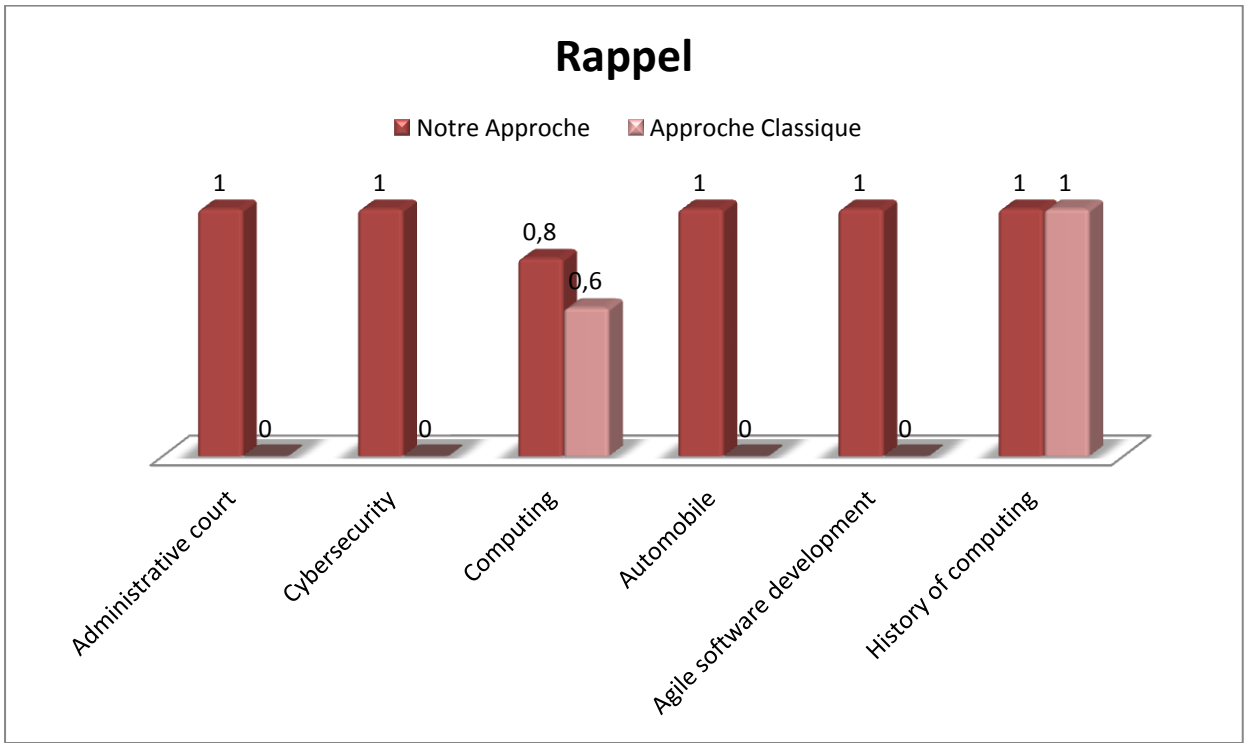


Figure IV.4 : Comparaison du Rappel des deux approches

Sur la majorité des requêtes, notre approche à donner des résultats similaires ou meilleurs par rapport à la Baseline, notamment pour les requêtes Q3 et Q6 celles-ci ont montré des résultats supérieurs à la Baseline.

	F-mesure	
Intitulé de la requête	**Notre Approche**	**Approche Classique**
Administrative court	**1**	**0**
Cybersecurity	**1**	**0**
Computing	**0,88 (88%) ▲**	**0,75 (75%) ▼**
Automobile	**1**	**0**
Agile software development	**1**	**0**
History of computing	**1 (99%) ▲**	**0,49 (49%) ▼**

Tableau IV.6 : F-mesure des deux approches

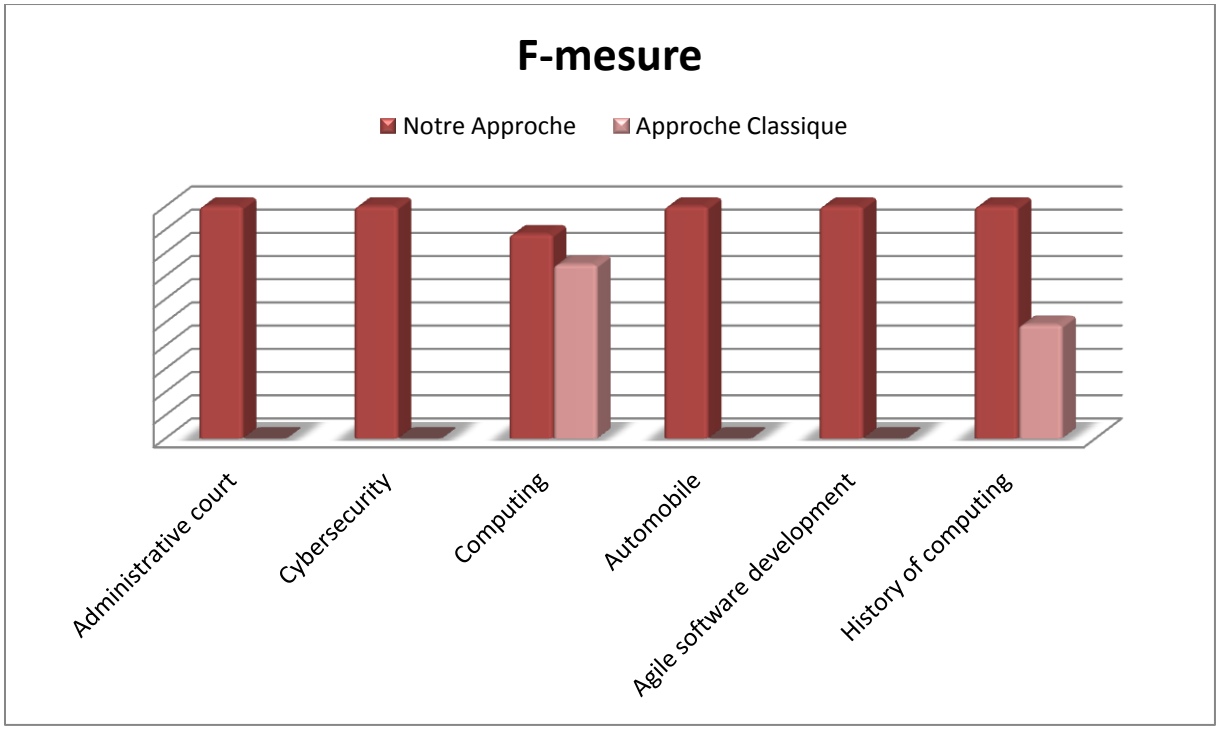


Figure IV.5 : Comparaison de la F-mesure entre les deux approches

Notre approche améliore nettement les résultats selon la métriques utilisée, ainsi nous avons obtenu une F-mesure supérieure à la Baseline, nous avons obtenu un taux moyen de 25%. La F-mesure n'étant pas la même pour les deux méthodes, nous concluons donc que notre approche augmente le nombre de documents pertinents restitués. En effet notre approche améliore le classement des documents pertinents par l'ajout de la sémantique à ces derniers.

IV.4.3. Synthèse

Après de nombreuses séries de test et d'expérimentations puis analyse des résultats, nous avons conclu que notre approche apportait de nettes améliorations en terme de sémantique par rapport à la méthode de recherche classique proposé par LUCENE, la collection sur laquelle nous avons effectuer nos expérimentations n'était pas volumineuse mais malgré cela, nous avons obtenu des résultats satisfaisants, cela démontre la pertinence de nos propositions et nous encouragent dans notre démarche.

IV.5. 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 la recherche classique. Nous sommes arrivés à la conclusion que notre approche apporte certaines améliorations en termes de sémantique 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 l'améliorer afin d'aboutir à de meilleurs résultats.

CONCLUSION GÈNÈRALE

CONCLUSION GÈNÈRALE

Les flux de microblogs sont devenus une ressource inestimable pour le marketing, la recherche, la diffusion de l'information ainsi que pour la gestion de la réputation en ligne (e-reputation management). La recherche et l'extraction de flux de microblog offres des défis intéressants et dans ce mémoire, nous avons présenté une méthode de liaison sémantique réussie pour les messages microblog. Les concepts identifiés, c'est à dire, les articles de Wikipedia, peuvent ensuite être utilisés pour, par exemple, l'exploitation minière des médias sociaux ou la présentation des résultats de recherche avancée. Notre nouvelle méthode est basée sur un classement de concept de haut rappel et une étape de sélection de concept de haute précision. En utilisant une collection d'essai construit à cet effet, nous avons montré qu'il surpasse nettement d'autres méthodes, y compris divers des approches récemment proposées.

Nous nous sommes concentrés principalement sur l'efficacité de la liaison sémantique dans le cadre de l'enrichissement des messages microblog.

Les travaux futurs comprennent les éléments suivants. Premièrement, bien que notre méthode ne soit dépendante de la langue en aucune façon, les annotations sémantiques sont spécifiques à la langue. d'autre part, Wikipedia contient déjà de nombreux liens inter-langues manuellement organisés que nous pourrions utiliser à cette fin. Deuxièmement, nous avons déjà mentionné une évaluation de notre méthode de liaison sémantique pour les travaux futurs on développera un Framework qui automatisera la collecte des nouveaux articles Wikipedia c.à.d. les nouveaux concepts qui enrichissent les messages microblogs. Nous reconnaissons également que notre échantillon de tweets, est comparativement petit (faible) et pourrait être pénalisant. Par conséquent nous avons l'intention d'appliquer les méthodes les plus performantes à un plus grand échantillon aléatoire de messages de microblog pour voir comment elles fonctionnent. En outre, dans ce mémoire, nous avons mis l'accent sur un domaine dépendent de façon à obtenir des classements de concept candidats de haut rappel. Nous croyons, cependant, que des informations supplémentaires, pourrait encore améliorer les performances de liaison sémantique.

Enfin, nous notons que Wikipedia contient quelques milliers de liens vers Twitter dans et nous avons l'intention d'examiner dans quelle mesure nous pouvons utiliser ces informations pour la liaison sémantique.

Bibliographie

Et

Référence

Bibliographie Et Références

[Abbas Nacira 14] «Vers une Extension Sémantique de l'Analyse Formelle de Concepts: Application à la Recherché d'Informations», Mémoire de Magister, Université Mouloud Mammeri Tizi Ouzou, 03 juillet 2014.

[Azzouz Wassila 13] « Contribution à la définition d'une approche d'indexation sématique de documents textuels» 2013. Mémoire de Magister, Université M'hamed Bougara Boumerdes 2005.

[**BAL 95**] Balpe J.-P., Lelu A., Saleh I., Hypertextes et hypermédias : Réalisations, outils et méthodes, Hermès, Paris, 1995.

[BAZIZ, 2005]. Mustapha BAZIZ. « Indexation Conceptuelle Guidée Par Ontologie Pour La Recherche D'information.» Thèse de doctorat, Université Paul Sabatier, 2005.

[Biemann, 05] Chris Biemann - Semantic Indexing with Typed Terms Using Rapid Annotation. in Methods and Applications of Semantic Indexing. Workshop at the 7th International Conference on Terminology and Knowledge Engineering. Copenhagen Denmark, Tuesday 16th August 2005.

[Belkin et al, 92] J.N Belkin, P. Ingwersen, A.M. « Proceedings of the 15 th annual International ACM SIGIR, In Conference on Research and Development in Information Retrieval». Copenhagen, Denmark, June, pages 21-24, ACM 1992.

[Bernstein et al, 10] Bernstein, M., Suh, B., Hong, L., Chen, J., Kairam, S., et Chi, E. (2010). Eddi : interactive topic-based browsing of social status streams. In Acm symposium on user interface software and technology (p. 303-312). New York, NY: ACM.

[CRO 92], Crouch C.J., Yang B., «Experiments in automatic statistical thesaurus construction», Proceedings of the ACM-SIGIR Conference on Research and development in information Retrieval, Copenhague, Danemark, p.77-88,1992.

[Damak 14] Damak Firas, « Etude des facteurs de pertinence dans la recherche de microblogs ». Recherche d'information [cs.IR]. Université Paul Sabatier, 2014. Français.

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

[DESMONTILS, JACQUIN, & MORIN, 2002] E. DESMONTILS, C. JACQUIN, E MORIN (2012). Indexation sémantique de documents sur le Web: application aux ressources humaines.

[Dong, Zhang, et al, 10] Dong, A., Zhang, R., Kolari, P., Bai, J., Diaz, F., Chang, Y., et al. (2010). Time is of the essence: improving recency ranking using twitter data. In In www

[Egozi et al, 11] O. Egozi, S. Markkovitch, E. Gabrilovich. « Based Information Retrieval using Explicicit Semantic Analysic ». ACM Transactions on Information Systems, Vol 29 Issue 2, April 2011.

[FOX 92], Fox C., « Lexical analysis and stoplis », p.102-130, dans Frakes W. B., Baeza-Yates R. (dir), Information Retrieval: Data, Structure and Algorithms, Prentice Hall, 1992.

[Krovetz et al., 92] R Korvetz, W. B. Croft. « Lexical Ambiguity and Information Retrieval».ACM Transactions on Information Systems, vol. 10, n°2, pages. 115-141. April 1992.

[Magnani et al, 12] ; Magnani, M., Montesi, D., et Rossi, L. (2012). Conversation retrieval for microblogging sites. Inf. Retr., 15 (3-4), 354-372.

[Mar 60], Maron M., Kuhns J., « On relevance, probabilistic indexing and information retrieval », Journal of the association for Computing Machinery, vol.7, p. 216-244, 1960.

[Massoudi et al, 11], Massoudi, K., Tsagkias, E., Rijke, M. de, et Weerkamp, W. (2011). Incorporating query expansion and quality indicators in searching microblog posts. In Ecir 2011: 33rd european conference on information retrieval (pp. 362–367). Dublin: Springer.

[Mihalcea et al., 00], R Mihalcea, D Moldovan. « Semantic indexing using WordNet senses ». In Proceedings of ACL workshop on IR and NLP, Hong Kong, October 2000.

[Missen et al, 09], Missen, M. M. S., Boughanem, M., et Cabanac, G. (2009, juin). Challenges for Sentence Level Opinion Detection in Blogs (regular paper). In International Conference on Computer and Information Science (ICIS), Shanghai, China, 01/06/2009-03/06/2009 (pp. 347–351). IEEE Computer Society.

[MOH 08] Bellot P., Boughanem M., « Recherche d'information et systèmes de questions-réponses», 2008 in " La recherche d'informations précises : traitement automatique de la langue, apprentissage et connaissances pour les systèmes de question-réponse (Traité IC2, série Informatique et systèmes d'information)", sous la direction de B.Grau, Hermès-Lavoisier, chapitre 1, p. 5

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

[Pang et Lee, 08], Pang, B., et Lee, L. (2008). Opinion mining and sentiment analysis. Found. Trends Inf. Retr., 2 (1-2), 1–135.

[R. Li et al, 12], Li, R., Lei, K. H., Khadiwala, R., et Chang, K.-C. (2012). Tedas : A twitter-based event detection and analysis system. In Data engineering (icde), 2012 ieee 28th international conference on (p. 1273-1276).

[Ramage et al, 10], Ramage, D., Dumais, S. T., et Liebling, D. J. (2010). Characterizing microblogs with topic models. In ICWSM'10 (pp. -1–1).

[Rijsbergen, 79] C.J. Van Rijsbergen «Information Retrieval », Butterworth-Heinemann, Newton,MA, USa, 2nd edition, 1979.

[Salton 75] G Salton, A Wong, CS Yang «A vector space model for automatic indexing», Communications of the ACM vol 18, n°11,p 613-620. November 1975.

[Salton 83] SALTON G., FOX E., W u H., «Extended Boolean information retrieval » Communications of the ACM, Vol. 31, n°2, p. 1002-1036, Novembre 1983.

[Sanderson, 94] Mark Sanderson. « Work sense disambiguation and information retrieval». In *Proceedings of* the 17th Annual International ACM_SIGIR Conference on Research and Development in information Retrieval, pages 142-151, Springer-Verlag, 1994.

[Song et al, 10] Song, S., Li, Q., et Zheng, N. (2010). A spatio-temporal framework for related topic search in micro-blogging. In Proceedings of the 6th international conference on active media technology (pp. 63–73). Berlin, Heidelberg : Springer-Verlag.

[Stairmand and Black, 1997] Mark A Staimand and WJ Black. « Conceptual and contextual indexing using wordnet-derived lexical chains». In *the proceedings of BCS IRSG Colloquium,* pages 74-65, 1997.

[Vosecky et al, 12] Vosecky, J., Leung, K. W.-T., et Ng, W. (2012). Searching for quality microblog posts : Filtering and ranking based on content analysis and implicit links. ,397-413.

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

[Yu et Hatzivassiloglou, 03], Towards Answering Opinion Questions: Separating Facts form Opinions and Identifying the Polarity of Opinion Sentences, Proceedings EMNLP-03, 8th CEM in Natural Language Processing, p.129-136, Sapporo, Japon, 2003.

[Wartik Steven 92], Boolean operations: Information Retrieval Data Structures and Algorithms.1992.

[Zhao et al, 11] Zhao, L., Zeng, Y., et Zhong, N. (2011). A weighted multi-factor algorithm for microblog search. In Proceedings of the 7th international conference on active media technology (pp. 153–161). Berlin, Heidelberg : Springer-Verlag.

[1] Edgar Meij, Wouter Weerkamp and Maarten de Rijke. Adding Semantics to Microblog Posts, 2012.

[2] D. Milne and I. H. Witten. Learning to link with Wikipedia. In CIKM ’08, 2008.

[3] P. Ferragina and U. Scaiella. Tagme: on-the-fly annotation of short text fragments (by wikipedia entities). In CIKM ’10, 2010.

[4] L. Breiman. Random forests. Mach. Learn., 45(1):5–32, 2001.

[5] K. W. Church and W. A. Gale. Inverse document frequency (IDF): A measure of deviations from poisson. In Proc. Third Workshop on Very Large Corpora, 1995.

[6] E. Meij, D. Trieschnigg, M. de Rijke, and W. Kraaij. Conceptual language models for domain-specific retrieval. Inf. Process. Manage., 46(4):448–469, 2010.

[7] E. Benson, A. Haghighi, and R. Barzilay. Event discovery in social media feeds. In ACL ’11, 2011.

Annexe

Annexe

1. 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.

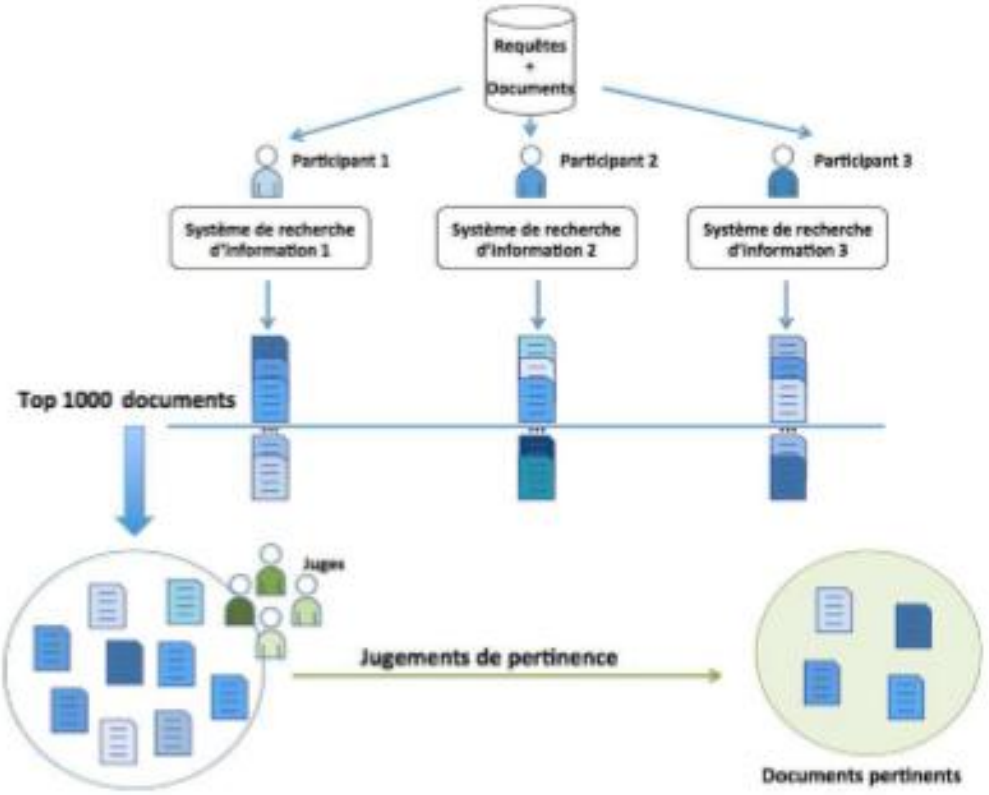


Figure A.1 : Campagne d'évaluation des SRI

2. La méthode de Voorhees

Voorhees (1993) [Voorhees, 93] a construit un outil de désambiguïsation basée sur le WordNet[1]. Pour désambiguïser une occurrence d'un mot ambigu, les synsets de ce mot sont classés en se basant sur la valeur de cooccurrence calculée entre le contexte de ce mot et un voisinage contenant les mots du synset dans la hiérarchie de WordNet. Voorhees a expérimenté cette approche sur une collection de test désambiguïsée (les requêtes de la collection de test sont aussi désambiguïsées manuellement) par rapport aux performances du même processus sur la même collection dans son état d'origine (ambigu).

Les résultats de ces expérimentations ont montré que pour chacune de ces collections, les performances du système de recherche d'informations diminuent sensiblement dans le cas de l'utilisation des collections désambiguïsées.

3. La méthode de Krovetz & Croft

Krovetz et Croft (1992) ont conduit une vaste étude sur certaines hypothèses ayant trait à la pertinence de la relation de correspondance du sens des mots dans la requête et les documents. En utilisant les collections de test CACM et Time, ils ont examiné les dix premiers documents restitués pour chaque requête. Leur but était de trouver l'existence d'une relation entre, d'une part, la correspondance/non-correspondance des sens et la pertinence / non-pertinence des documents restitués, d'autre part. Leurs résultats ont montré que la relation de non-correspondance de sens a tendance à apparaître de façon significative dans les documents non pertinents. De plus, ils ont remarqué que la non-correspondance de sens est plus rare dans les documents classés en premier. Krovetz et Croft expliquent ceci par deux raisons : L'effet de la collocation des mots des requêtes qui fait qu'ils contribuent ensemble à exprimer un même sens, même si ils sont isolément ambigus. Et la distribution non uniforme du sens des mots dans les documents. Krovetz et Croft ont montré que les différents sens d'un mot n'occurrent pas dans une même proportion et que souvent un sens d'un mot (le sens dominant) occurrent plus fréquemment que les autres sens.

4. Le modèle OntoSeek

Le système OntoSeek de Guarino et al (1999) est l'une des premières tentatives expérimentales sur l'application des ontologies à la recherche d'informations ; il est considéré comme une excellente référence dans ce domaine. Il est conçu pour traiter les pages jaunes et les catalogues de produits (en ligne). Le contenu des pages ainsi que les requêtes sont modélisés par un formalisme basique de graphes conceptuels. OntoSeek combine un

1. WordNet : WordNet est un réseau sémantique organisé autour de la notion de synset (Synonyme set : ensemble de synonymes). Un synset regroupe des termes (simples ou composés) ayant un même sens dans un contexte donné. Les synsets sont liés par différentes relations telles que l'Hyperonymie (is-a) et son inverse, l'Hyponymie (instance-de).

appariement basé sur le contenu et guidé par ontologie (ontology-driven content-matching). Notons que dans ce système, le processus de désambiguïsation se fait de manière interactive.

5. Le modèle DocCore

Dans ce modèle présenté dans [BAZIZ, 2005], la représentation du document, appelée noyau sémantique, est une représentation analogue à celle du réseau sémantique à la différence que les arcs ne sont pas étiquetés, mais quantifiés par une valeur réelle dénotant la proximité sémantique entre les deux concepts. Cette valeur de proximité sémantique est calculée en utilisant des mesures de similarité sémantique.
Le choix des valeurs au détriment des étiquettes permet la désambiguïsation des termes du document, ainsi que l'assignation d'un « poids sémantique » aux différents nœuds du réseau.
L'approche consiste à projeter les documents sur une ontologie linguistique générale, telle que WordNet. Il s'agit d'identifier pour chaque document les représentants de concepts de l'ontologie. Ces derniers peuvent être des mots simples ou des groupes de mots

6. PageRank

PR est l'algorithme d'analyse des liens concourant au système de classement des pages Web utilisé par le moteur de recherche Google. Il mesure quantitativement la popularité d'une page web. Le PageRank n'est qu'un indicateur parmi d'autres dans l'algorithme qui permet de classer les pages du Web dans les résultats de recherche de Google. Ce système a été inventé par Larry Page, cofondateur de Google.

7. Le modèle de sac de mots

Le modèle sac de mots est une représentation simplifiée utilisé dans le traitement du langage naturel et de récupération d'informations (IR). Dans ce modèle, un texte (comme une phrase ou un document) est représenté comme le sac de ses mots, sans tenir compte de la grammaire et de l'ordre des mots.

8. Wikipedia

Wikipedia est une encyclopédie en ligne universelle et multilingue. Elle est en cours de rédaction collaborative sur Internet avec la technologie wiki. Wikipedia a pour principe d'offrir un contenu libre, neutre et vérifiable.

Un wiki est un système de gestion de contenu de site Web qui rend les pages Web librement et également modifiables par tous les visiteurs autorisés. On utilise les wikis pour faciliter l'écriture collaborative de documents avec un minimum de contraintes. Le wiki a été inventé en 1995 par Ward Cunningham, pour une section d'un site sur la programmation informatique qu'il a appelée WikiWikiWeb. Le mot « wiki » vient du redoublement hawaiien wiki wiki, qui signifie « rapide ». L'encyclopédie Wikipedia est devenue le plus visité des sites Web écrits avec un wiki. (Source : http://fr.wikipedia.org/wiki/Wiki)

9. DBpedia

DBpedia est un projet universitaire et communautaire d'exploration et extraction automatiques de données dérivées de Wikipedia. Son principe est de proposer une version structurée et sous forme de données normalisées au format du web sémantique des contenus encyclopédiques de chaque fiche encyclopédique. DBpedia vise aussi à relier à Wikipédia (et inversement) des ensembles d'autres données ouvertes provenant du Web des données. Ce projet est conduit par l'université de Leipzig, l'université libre de Berlin et l'entreprise OpenLink Software.

10.Le Web des données

(linked data, en anglais) est une initiative du W3C (Consortium World Wide Web) visant à favoriser la publication de données structurées sur le Web, non pas sous la forme de silos de données isolés les uns des autres, mais en les reliant entre elles pour constituer un réseau global d'informations.

11.DBpedia-Spotlight

DBpedia Spotlight est un outil d'annotation sémantique de texte utilisant les ressources de DBpedia en tant que liens documentaires, fournissant une solution pour relier les sources d'information non structurées à DBpedia. DBpedia Spotlight reconnaît que les noms des concepts ou entités ont été mentionnés (par exemple, "Michael Jordan"), et par la suite correspond à ces noms à des identificateurs uniques (par exemple dbpedia: Michael_I. _Jordan , le professeur d'apprentissage automatique ou dbpedia: Michael_Jordan le joueur de basket - ball). Il peut également être utilisé pour la construction de votre solution pour les tâches d'extraction d'information.

DBpedia Spotlight offre plusieurs façons de l'utilisé, doit comme Web Service, ou bien comme JAR avec toutes les dépendances inclus qu'on peut implémenter et exécuter dans un

IDE. (Source : //github.com/dbpedia-spotlight/dbpedia-spotlight/wiki/Introduction). Cette dernière est la manière dont la quelle nous l'avons introduite dans notre projet (Figure A.1) :

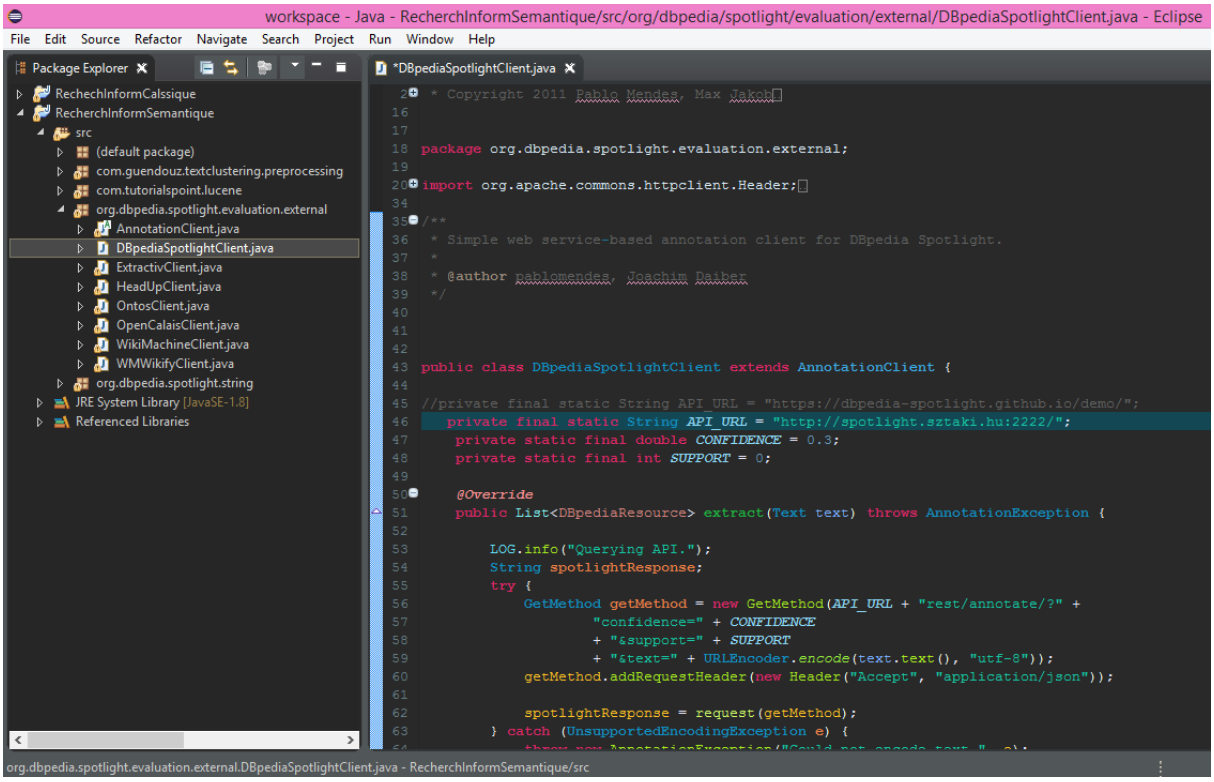


Figure A.2 : Classe java DBpedia_Spotlight_Client

La classe en dessus (Figure A.1) est la classe principale de l'API elle fonctionne selon l'enchainement suivant :

Repérage (Spotting) : prend le texte en entrée et reconnaît les entités / concepts pour annoter. Plusieurs techniques de repérage sont disponibles, tels que le dictionnaire recherche et reconnaissance d'entités nommées (NER).

Désambigüisation : prend la saisie du texte, où les entités / concepts ont déjà été reconnus et Choisit un identifiant pour chaque entité / concept reconnu étant donné le contexte.

Annotation : Prend le texte (résultant de l'étape précédente) comme entrée, reconnaît les entités / concepts pour annoter et choisit un identifiant pour chaque entité / concept reconnu étant donné le contexte. Enfin nous retourne une liste des concepts clés.

12.Annotation sémantique

L'annotation est le processus qui consiste à attacher des informations complémentaires au contenu textuel d'un document. L'annotation sémantique consiste à relier ces contenus à des informations précises en relation avec l'identité sémantique des données annotées. On

peut ainsi définir l'annotation sémantique comme la tâche permettant de déterminer l'identité exacte d'un concept contenu dans un texte et de fournir des informations sur ce concept.

13.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.

14.Lucene

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

• Classes d'indexation :

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 : 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 qu'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.

15.Eclipse

Eclipse est un projet, décliné et organisé en un ensemble de sous-projets de développements logiciels, de la fondation Eclipse visant à développer un environnement de production de logiciels libre qui soit extensible, universel et polyvalent, en s'appuyant principalement sur Java.

Son objectif est de produire et fournir des outils pour la réalisation de logiciels, englobant les activités de programmation, Bien qu'Eclipse ait d'abord été conçu uniquement pour produire des environnements de développement, les utilisateurs et contributeurs se sont rapidement mis à réutiliser ses briques logicielles pour des applications clientes classiques. Cela a conduit à une extension du périmètre initial d'Eclipse à toute production de logiciel : c'est l'apparition du framework Eclipse RCP en 2004.

Figurant parmi les grandes réussites de l'Open source, Eclipse est devenu un standard du marché des logiciels de développement, intégré par de grands éditeurs logiciels et sociétés de services.

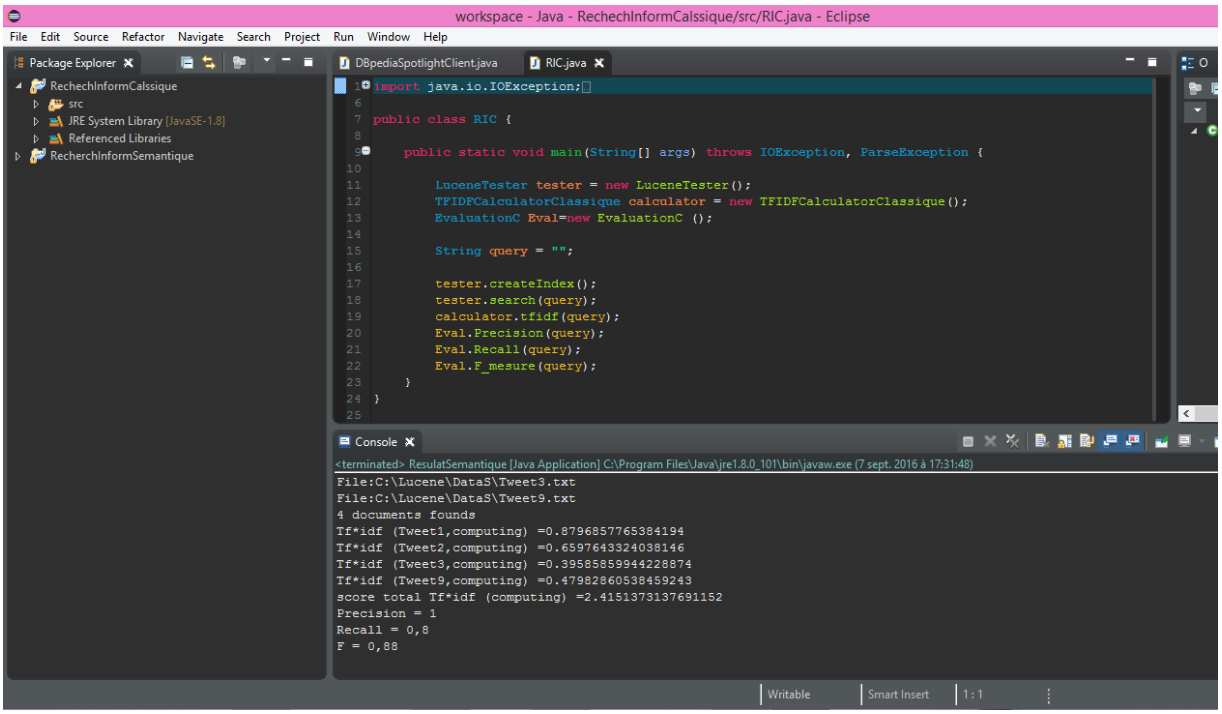


Figure A.3 : Interface de travail Eclipse

16. Pear Analytics

Pear Analytics est une agence de marketing numérique avec un accent particulier sur l'analyse, la mesure, l'optimisation de la conversion, génération de prospects et de marketing de contenu.

17.Twitter Curator

Curator est, comme son nom l'indique, un outil de curation permettant de sélectionner des tweets.Il s'articule avant tout sur la recherche de contenus sur Twitter. Cette recherche peut s'appuyer sur un puissant outil de filtrage permettant de sélectionner avec une très grande précision le contenu recherché.

18.Notre collection de tweets:

Tweet 1

jean cornelis (@JeanCrnls)	16-Aug-2016 08:52
Our expert's team in computing, will indulge to realize your project, by a well-kept work https://t.co/X7T7JrMmle	

Tweet 2

Simon Denier (@simondenier)	27-Aug-2016 08:46
A little of history of the computing and the networks: Louis Pouzin https://t.co/D2Jm6XoXco via @inria	

Tweet 3

Camille Polloni (@CamillePolloni)	27-Aug-2016 08:24
Perquis. admin. in Roubaix: against the advice of AC, the Council of State confirms the exploitation of the computing data https://t.co/0sFffwDTcc	

Tweet 4

Cdiscount (@Cdiscount)	27-Aug-2016 08:01

Comeback in rock-bottom prices! LAPTOP PC + MOUSE = the assured good grades https://t.co/uqLmQRsKsO https://t.co/ILTL2THPQK	

Tweet 5

Alexandre Cayuela ♞ (@alexandrejimen)	23-Aug-2016 09:10
Scrum on the grill of the jurisprudence https://t.co/Ue3wPU3sDm via @LUsineDigitale	

Tweet 6

MahreiAzadi (@AzadiMahrei)	27-Aug-2016 09:14
Mrs RoyalSegolene do you intend to visit the cemetery of Khavaran where rest the bodies of prisoners politics executed ? #lemontsaintmichel	

Tweet 7

Maurice Leroy (@MauriceLeroy)	27-Aug-2016 09:14
To read this morning in it @NRBlois: " Maurice Leroy, the centrist asset of Nicolas Sarkozy» https://t.co/3ZmlOcKXTp	

Tweet 8

libertemaloya (@libertemaloya)	27-Aug-2016 09:13
Gaza: the UPR asks to the French government to give some explanation about "double standards" of her foreign policy and @UPR_Asselineau	

Tweet 9

Kaspersky Lab France (@kasperskyfrance)	16-Aug-2016 08:57
Everything you need to know about computer security and threats ! https://t.co/L4rLiMYnul	

Tweet 10

buy fine products (@ProductsFine)	29-Aug-2016 10:32
2010 HOT WHEELS HW PREMIERE TOYOTA LAND CRUISER FJ40 YELLOW BLUE LONG CARD https://t.co/SbT3IIpfyR https://t.co/KPhR5voh0o