Université Abou Bekr Belkaid
Tlemcen Algérie
جامعة أبي بكر بلقايد
تلمسان الجزائر

République Algérienne Démocratique et Populaire
Université Abou Bakr Belkaid– Tlemcen
Faculté des Sciences
Département d'Informatique

Mémoire de fin d'études

pour l'obtention du diplôme de Master en Informatique

Option: *Système d'Information et de Connaissances (S.I.C)*

Thème

Conception et implémentation d'un système de recherche à base d'annotations sociales

Réalisé par :

- MEHIDI Tawfiq
- RABAH Zakarya

Présenté le 24 Juin 2014 devant le jury composé de MM.

- Messabihi M. *(Président)*
- Hadjila F. *(Encadreur)*
- Khitri S. *(Examinateur)*
- *Kazi A.* *(Examinateur)*

Année universitaire: 2013-2014

REMERCIEMENT

Nous remercions tout d'abord le bon Dieu, le tout puissant de nous avoir armé de force et de courage pour mener à terme ce projet.

Nous tenons à remercier particulièrement notre encadreur Mr Hadjila pour ses conseils fructueux et pour son aide précieuse qui nous a conduits à concrétiser ce travail.

Nous remercierons très sincèrement, les membres de jury d'avoir bien voulu accepter de faire partie de la commission d'examinateur.

Merci à tous ceux qui ont contribué de près ou de loin à ce que la réalisation de ce projet soit possible.

Nous tenons également à remercier tous nos collègues de promotion que nous avons eu le plaisir de les côtoyer pendant les années d'étude.

Nous remercierons tous ceux qui ont contribué de près ou de loin à la réalisation de ce travail.

DÉDICACES

Toutes les lettres ne sauraient trouver les mots qu'il faut ...
Tous les mots ne sauraient exprimer la gratitude, l'amour, le respect, la reconnaissance ...
Aussi, c'est tout simplement que ...
Ce travail représente l'aboutissement du soutien et des encouragements que
mes parents m'ont prodigués tout au long de ma scolarité.
Je dédie humblement ce manuscrit à...?

A mon très cher père...
Aucune dédicace ne saurait exprimer à sa juste valeur tout l'amour, le respect, l'attachement et la reconnaissance que je te porte.
Tu m'as enseigné la droiture, le respect et la conscience du devoir.
Ce travail est le fruit de tous tes sacrifices, tes encouragements, ton désir de me voir arriver et ton soutien permanent durant ce long parcours.
Puisse Dieu, le tout puissant, te procurer santé, bonheur et longue vie...

A ma très chère et douce mère...
A la plus merveilleuse des mères.
J'espère réaliser, en ce jour, l'un de tes rêves.
Aucun mot ne saurait exprimer mon respect, ma considération et l'amour que je te porte.
Ta présence constante à mes cotés, tes encouragements et tes prières m'ont été d'une aide précieuse et m'ont permis d'atteindre le but désiré.
Puisse Dieu le tout puissant te donner santé et longue vie afin que je puisse te combler à mon tour...

A mon très cher frère

ABDELHAK

En témoignage de l'attachement, de l'amour et de l'affection que je porte pour toi. Je te dédie ce travail avec tous mes vœux de bonheur, de santé et de réussite.

A mes très chères sœurs… ainsi que leurs *enfants*
DOUAA et BENALI

A
Ma Chère grand-mère paternelle
Ma Chère grand-mère maternelle
Que ce modeste travail, soit l'expression des vœux que vous n'avez cessé de formuler dans vos prières.

Que Dieu vous préservé sante et longue vie.

A mon binôme Zaki et à toute sa famille

A mes chers oncles, tantes, cousins (es)Que ce travail soit le témoignage de ma grande affection et mes sentiments les plus sincères.
Je prie DIEU de vous réserver le bonheur et, la santé.

A toute ma famille,
Avec toute mon affection et mon respect.

A tous mes ami(e)s et tous ceux qui me sont chers.

A eux tous, je souhaite un avenir plein de joie, de bonheur et de succès.

A tous ceux qui ont contribué de loin ou de près à l'élaboration de ce travail,

Avec tous mes remerciements.

A tous mes professeurs et maîtres,
Avec tous mes respects et mon éternelle reconnaissance.

A toutes la promotion de Master SIC : 2013-2014
A tous ceux que j'ai omis de citer En témoignage sincère d'affection et de nobles sentiments

MEHIDI Tawfiq

DÉDICACES

Je commence par rendre grâce à dieu et à sa bonté, pour la patience, la compétence et le courage qu'il m'a donné pour arriver à ce stade.

A mes très chers parents qui ont toujours été là pour moi, et qui m'ont donné un magnifique modèle de labeur et de persévérance. J'espère qu'ils trouveront dans ce travail toute ma reconnaissance et tout mon amour.

A mes chers frères et sœurs et leurs enfants

A mes tantes et à mes oncles.

A chaque cousins et cousines.

A tous mes amis

A Mon binôme Tawfiq
Avec qui j'ai partagé les joies et les difficultés relatives au suivi de ce projet, pour sa motivation et ses judicieuses propositions.

A toute personne qui a contribué de près ou de loin à la réalisation de ce travail.

A toutes la promotion de Master SIC : 2013-2014

RABAH Zakarya

Résumé

L'explosion du Web 2.0 (blogs, wikis, sites de partage, réseaux sociaux, etc.) entraine de nouveaux modes de navigation et de recherche d'information sur Internet. L'information n'est plus uniquement recherchée sur des moteurs de recherche tels que Google ou Bing, elle est désormais filtrée via les réseaux sociaux (Flickr, Delicious...), ou découverte au hasard de la navigation. Notre travail s'inscrit dans le domaine de la Recherche d'Information (RI). Il a pour objet la création d'un modèle de recherche utilisant les réseaux sociaux pour intégrer les annotations utilisateur de chaque document dans le calcul de similarités entre les documents textuels et la requête d'utilisateur en vue d'améliorer le processus de recherche d'information. Dans ce système, la pertinence d'un document est estimée par la combinaison de la pertinence thématique (doc-requête) et de la pertinence sociale (doc-tags).

Mots clés : SRI, web social, mesures de similarité, ontologies, annotations.

Abstract

The explosion of the web 2.0 (blogs , wikis, sharing sites, social networks , etc ..) leads causes new modes of navigations and research of information on internet . Information is no more researched in search engines like google or bing only, it is now filtered through social networks (Flickr, Delicious…) or randomly discovered in navigation. Our work is in the field of Information Retrieval (I.R). It aims to create a model of research using the social networks for incorporate user annotations of every document in similarity calculations between textual documents and user request to improve the process of Information Retrieval. In this system, the relevance of a document is estimated by the combination of topical relevance (doc-query) and of social relevance (doc-tags).

Keywords: (RSI, social web, similarity measures, ontology's, annotations).

ملخص

أدى ظهور الجيل الثاني من الويب (ويب 2.0) إلى انبثاق تقنيات حديثة (المدونات,الو سوم ,تقنية الويكي, والشبكات الاجتماعية,الخ...) والتي بدورها أدت إلى انبثاق أنماط جديدة من طرق التصفح والبحث عن المعلومات على الانترنت.إن نظام استرجاع المعلومات لم يعد قائما على محركات البحث العادية مثل غووغل و بينغ فقط بل ذهب إلى ابعد الحدود و ذلك بإدراج الشبكات الاجتماعية في البحث (فليكر,ديليشوس).

عملنا هذا يندرج في مجال استرجاع المعومات والذي يهدف إلى خلق نموذج للبحث باستخدام الشبكات الاجتماعية كمساعد وذلك بدمج شروح وتعليقات المستخدم لكل وثيقة في حساب التشابه بين نصوص الوثائق و طلب المستخدم من أجل تحسين عملية البحث عن المعلومات.

في هذا النظام أهمية الوثيقة هي مقدرة عن طريق الجمع بين الأهمية الموضوعية و الأهمية الاجتماعية.

الكلمات المفتاحية : نظام البحث عن المعلومات، انطولوجيا،الشبكة الاجتماعية ، شروحات.

.

Table de matières

Introduction générale

Contexte

Notre travail se place dans le domaine de la recherche d'information. Il est intitulé "conception et implémentation d'un système de recherche d'information à base d'annotations sociales"

La RI est un domaine qui s'est apparue depuis les années 1950-1960 [Yaël, 2009], quelque années après l'invention de l'ordinateur. Ce domaine est né pour automatiser la RI dans les bibliothèques. Elle s'intéresse à l'acquisition, l'organisation, stockage et la recherche de l'information.

L'objectif de la RI c'est de créer et développer un SRI pour faciliter l'accès à une collection de documents informatisées (corpus) à un besoin en information de l'utilisateur exprimé sous forme d'une requête c'est-à-dire sélectionner les documents qui sont pertinent pour lui.

Problématique et contribution

Tout le monde sait que l'accès et l'utilisation de la recherche d'information toute seule n'est pas suffisante.

Tout système de recherche d'informations (SRI) a pour but de satisfaire les besoins des utilisateurs mais il arrive, malgré toutes les techniques intégrées dans ces systèmes, que les utilisateurs n'obtiennent pas des résultats pertinents.

Ainsi, les nombreux chercheurs intègrent plusieurs astuces pour améliorer les résultats de la recherche des documents pertinentes. Parmi ceux-ci "retour d'information"(feedback).

Dans ce travail nous essayons de répondre à cette question à travers l'intégration de « les retours utilisateurs » ou (feedback user) Dans les systèmes de recherche d'information en prenant en compte les remarques et les besoins de l'utilisateur sur les documents qui devra être comprise aussi bien par l'humain, que par la machine.

Plan de travail

Ce mémoire est organisé en trois chapitres :

Le chapitre 1 intitulé **La recherche de l'information** traite des généralités concernant le domaine de la recherche d'information présente les notions et les concepts de base de la RI ainsi que les principaux modèles de existants. Enfin nous avons expliqué la technique d'évaluation d'un système de recherche d'information, aussi une petite vision sur les deux

sites sociaux Flickr et Delicious.

Le chapitre 2 intitulé **Les générations du Web** est consacré pour les différentes générations du Web, en premier lieu le Web 0 ou le Web militaire, en deuxième lieu le Web 1.0 ou le web traditionnel appelé aussi le Web statique, en troisième lieu le Web 2.0 ou le Web participatif, puis le Web 3.0 (le Web sémantique), et en dernier lieu c'est le Web 4.0 qui a plusieurs appellations parmi eux le Web of thing.

La dernière partie nommé **Conception et implémentation du prototype**, concerne la présentation de notre approche qui consiste mettre en œuvre un système de recherche d'information prenant en compte les annotations issus de tous les utilisateurs. Au final, nous dressons un bilan de nos travaux et nous présentons ensuite les perspectives d'évolution de ces travaux.

Chapitre I :
La Recherche d'information

I. Introduction

Le volume de l'information disponible électroniquement (texte, son, vidéo, image…) est toujours plus importants et par conséquent la recherche des documents pertinents devient de plus en plus difficile [Gilles, 2010].

La RI est apparue comme une réponse au besoin de gérer la quantité d'information. Elle est historiquement liée aux sciences de recherche [Yaël, 2009] qui intègre des méthodes et des algorithmes pour faciliter l'accès à l'information pertinents [Ressad-Bouidghaen, 2011] autrement dit pour satisfaire le besoin de l'utilisateur écrit sous forme d'une requête en langage naturel composée des mots clés.

Un SRI est un ensemble de logiciels assurant l'ensemble des fonctions nécessaires à la recherche de l'information. L'objectif principal de ce système est de mettre en œuvre un processus de comparaison (modèles et algorithmes) entre la requête et une collection de documents dont le but ceux qui sont pertinents.

II. Bref Historique de la RI

-***1940:*** Apparition des SRI, focalisation de la RI sur les applications dans des bibliothèques. D'où aussi le nom "automatisation de bibliothèques".

-***1950:*** Apparition du modèle booléen et l'élaboration de petites expérimentations sur des petites collections de documents (références bibliographiques).

-***1960 et 1970:*** Apparition du système SMART, Développement d'une méthodologie d'évaluation de système et conception de corpus de test(CACM).

-***1980:*** Développement de l'intelligence artificielle, ainsi on tentait d'intégrer des techniques de l'IA en RI (système expert).

-***1990 et 1995:*** L'apparition d'internet, la RI a été modifié et sa problématique plus élargie.

III. Les concepts fondamentaux de la recherche d'information

III.1 Notions de base

III.1.1 RI

« La Recherche d'Information (RI) est une branche de l'informatique qui s'intéresse à l'acquisition, l'organisation, le stockage, la recherche et la sélection d'information. » [Salton & al, 1989]

« La recherche d'information est la science qui étudie la manière de répondre pertinemment à une requête en retrouvant de l'information dans un corpus. » [Christopher & al, 2009]

« Ensemble des méthodes, procédures et techniques permettant, en fonction de critères de recherche propres à l'usager, de sélectionner l'information dans un ou plusieurs fonds de documents plus ou moins structurés.»[Le Targat, 2005]

III.1.2 SRI

Un SRI peut être défini comme l'ensemble des procédures et des opérations permettant la gestion, la représentation, l'interrogation, la recherche, le stockage et la sélection des informations répondant aux besoins d'un utilisateur [Brini, 2005].

III.1.3 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 être un texte, un morceau de texte, une image, une bande vidéo, etc. [Daoud, 2009]

III.1.4 Requête

Une requête est une formulation du besoin d'information d'un utilisateur. Elle peut être vue comme étant une description sommaire des documents ciblés par la recherche. Pour une recherche documentaire donnée, l'utilisateur doit soumettre une requête au moteur de recherche dans laquelle il spécifie les mots clés représentant son besoin en information. [Daoud, 2009]

III.1.5 Pertinence

Selon les premières définitions de la pertinence, la pertinence est la correspondance entre un document et une requête ; une mesure de l'informativité du document à la requête ; un degré de relation (chevauchement, etc.) entre le document et la requête ; etc. La notion de pertinence est le critère primaire pour l'évaluation des systèmes de recherche d'informations. Le processus de jugement de la pertinence de l'information est basé sur le degré de similitude de la représentation de la requête avec le contenu du document retrouvé par le système. [Daoud, 2009]

III.2 Processus générale de RI

Un SRI génère un processus qui a pour rôle de retrouver le maximum des documents pertinent en comparant le besoin en information de l'utilisateur exprimé sous forme d'une requête avec la base documentaire disponible.

La pertinence dépend de l'utilisateur, c'est-à-dire difficile à automatiser, on parle souvent de deux types de pertinence :

– **la pertinence utilisateur** : le document est jugé pertinent par l'utilisateur en fonction de son besoin en information.

– **la pertinence système** : le document est juge pertinent par le SRI pour une requête sur la base de la fonction de pertinence. [Nassr, 2002]

Le processus de recherche appelé actuellement « le processus en **U** de la RI » illustré dans la figure I.1**[**Marini, 2010] :

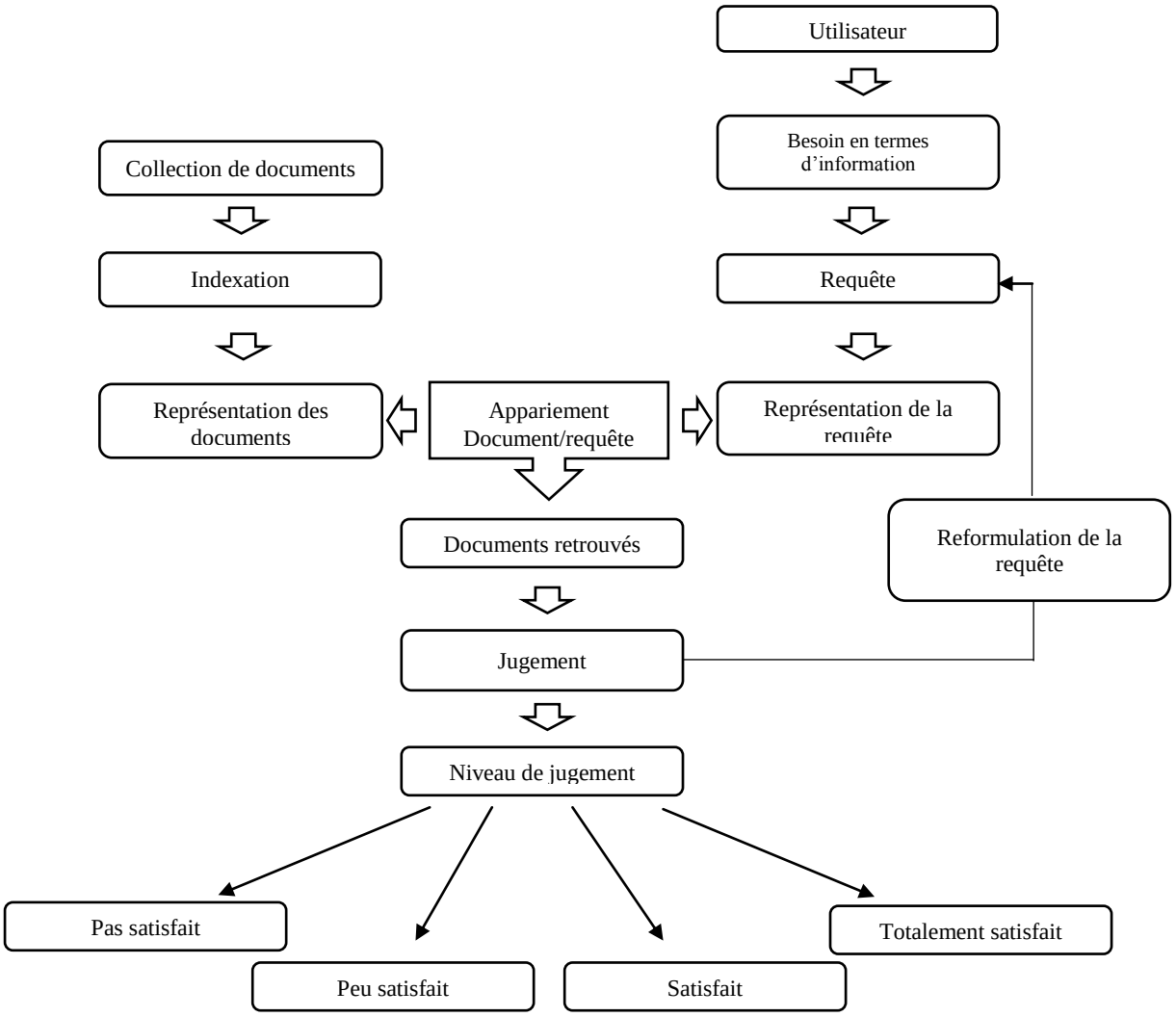

Figure I.1: processus de recherche d'information

Ce dernier est décomposé en trois fonctions :

- L'indexation
- L'appariement document/requête
- Reformulation de la requête

III.2.1 Processus d'indexation

La recherche d'information(RI) par un parcours complet de tous les textes de la collection de documents n'est pas pratique [Ressad-Bouidghaen, 2011]. Donc on utilise une opération appelée l'indexation qui consiste à créer un ensemble des mots clés (termes) à partir de l'analyse d'un document qui se trouve dans une collection de documents pour que l'exploitation de ces mots clés ou descripteur par le système soit facile. Ces mots clés peuvent être regroupés dans un thésaurus (« en pratique, un

thesaurus regroupe plusieurs relations de types linguistique (équivalence, association, hiérarchie) et statistique (pondération) ») [Baziz, 2005].
L'objectif de cette opération est de garder les termes significatifs de ce document.

III.2.2 Les types d'indexation

III.2.2.1 Manuel

Chaque document est analysé par un documentaliste ou par un spécialiste du domaine, qui extrait les mots basant sur un vocabulaire contrôlé (liste hiérarchique, thesaurus, lexique,. . .) [Ressad-Bouidghaen, 2011]. L'avantage de l'indexation manuelle est d'assurer un meilleur rapport entre les documents et les termes choisis par les spécialistes [Baziz, 2005].mais l'inconvénient elle est couteuse en terme de temps (nécessite un temps important) et plus d'effort intellectuel (nombres de personnes).

III.2.2.2 Automatique

L'indexation dans ce cas-là est faite par un SRI basé sur des algorithmes et des méthodes, l'expert du domaine n'intervient pas [Daoud, 2009]. Elle détecte d'une façon automatique les concepts significatifs d'un document en analysant le document mot par mot aussi l'élimination des mots vides, la lemmatisation, la pondération des termes, à la fin la création de l'index. Ce type d'indexation est souvent le plus utilisé. [Nassr, 2002]

III.2.2.3 Semi-automatique

Appelée aussi indexation supervisée. Ce type d'indexation fait une combinaison des deux modes précédant [Daoud, 2009], les termes du document sont extraits en un premier temps par un processus automatique. Mais le choix final des mots clés est fait par l'indexeur ou le spécialiste du domaine, généralement les indexeurs utilisent un vocabulaire contrôlé sous forme de thesaurus ou de base terminologique [Baziz, 2005].

III.2.3 Les étapes de l'indexation :

Le processus de l'indexation passe par des étapes afin que l'index soit créé comme il est illustré dans la figure I.2 :

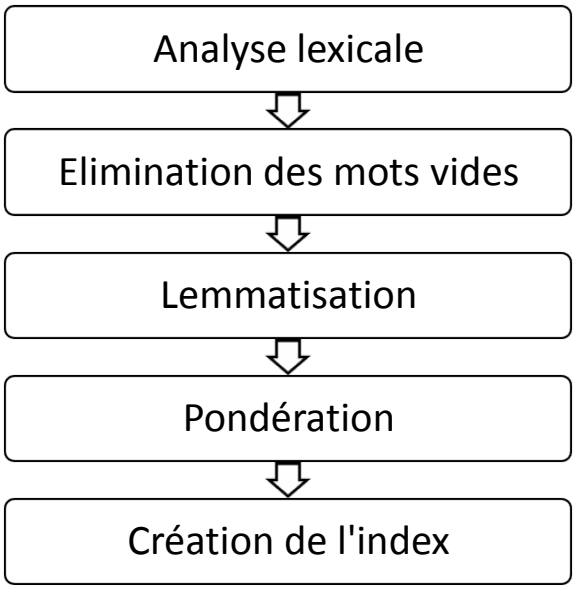

Figure I.2: Les étapes de l'indexation

III.2.3.1 L'analyse lexicale :

C'est l'opération de transformer un document textuel en un ensemble de termes ou unité lexicale en reconnaissant les espaces de s´séparation des mots, des caractères spéciaux, des chiffres, les ponctuations ou une liste de séparateurs. [Ho, 2004]
Il y a des cas particuliers où les mots sont composés ou contiennent des séparateurs

Ex : aujourd'hui, pomme de terre, Le Mans, 127.0.0.1, M. Durand, 14/07/1789.

III.2.3.2 Elimination des mots vides

C'est la suppression des mots de fort fréquence ou à faible contenu informatif ces mots appelé mots vides (tels que les pronoms personnels, les articles, les mots de liaison, ou les prépositions), pour ne garder que les termes importants. Plusieurs techniques peuvent être mises en œuvre parmi celles-ci, l'utilisation des stops liste ou des anti-dictionnaires (anti-lexiques) et l'utilisation des mesures statistiques. Le traitement lié à un anti-dictionnaire est très simple si un mot est apparait dans l'anti-dictionnaire et dans

le texte a indexé, il n'est pas considéré comme un index, il peut cependant induire des effets de silence (par exemple, en éliminant le mot **a** de **vitamine a** [Kompaoré, 2008].

III.2.3.3 Lemmatisation (radicalisation)

C'est une étape qui a pour but de regrouper les diffèrent variantes d'un mot a sa forme canonique ou lemme dans l'exemple de [Yaël, 2009] « il peut être utile de retrouver des documents contenant les mots « transmission », « transmis », « transmet », « transmettra », «transmetteur» à partir d'une requête comportant le mot « transmettre ». Pour cela il est possible d'éliminer les différences non significatives et de garder la partie commune. Sur l'exemple, les mots ont la même racine (le lemme) et une terminaison différente ».
Pour les verbes conjugués il suffit de rendre le verbe à l'infinitif pour les conjugués et le singulier pour les noms. Des fois le passage à la forme canonique supprime le sens du mot comme le verbe portera et le nom porte seront indexés de la même façon.
Parmi les techniques utilisées dans radicalisation les suivantes :

- La table de consultation (dictionnaire).
- L'élimination des affixes (Porter).
- La troncature.
- Les variétés de successeurs (n-grammes).
- L'utilisation des étiqueteurs grammaticaux (taggeurs).

III.2.3.4 Pondération des termes

Le poids d'un terme dans un document traduit l'importance de ce terme dans le document. La pondération est une fonction fondamentale, elle est généralement basé sur l'association des valeurs numériques aux termes de manière à représenter le pouvoir de discrimination (degré d'informativité) de ces termes pour chaque document de la collection, L'objectif est de trouver les termes qui représentent le mieux le contenu d'un document.
La pondération d'un terme s'exprime en fonction de deux pondérations :

- Une pondération locale reflète l'importance locale du terme dans le document (mesures statistiques locales).

- La pondération globale exprime l'importance globale du terme dans la collection (mesures statistiques globale).

La majorité de ces mesures tirent leur origine de la loi de Zipf et de la conjecture de Kuhn [Harrathi, 2010].

III.2.4 Quelques méthodes de pondération

III.2.4.1 Loi de Zipf

La loi de Zipf,[1] également appelée loi rang-fréquence, c'est une loi empirique .Cette loi est mise en évidence par le linguiste américain George K. Zipf (1949) [Caron, 2004]. Dans [Yaël, 2009] l'auteur a dit que « la fréquence d'un mot est inversement proportionnelle à son rang dans la liste des termes classés par fréquence décroissante ou encore que le produit de la fréquence de n'importe quel mot par son rang est constant ». La relation est donnée par la formule suivante :

$$\text{Rang} \times \text{fréquence} = \text{constante}$$

Le rang d'un mot est sa position dans cette liste [Baziz, 2005].

Donc l'utilisation de loi de Zipf dans le domaine de (RI) c'est pour déterminer les mots qui représentant au mieux le contenu d'un document, par l'élimination des termes trop fréquents ou trop rares.

Le rapport entre le rang × fréquence et importance du terme est représenter par les courbes suivant (Figure I.3) [Yaël, 2009] :

[1] http://www.lewishistoricalsociety.com/wiki/tiki-read_article.php?articleId=75

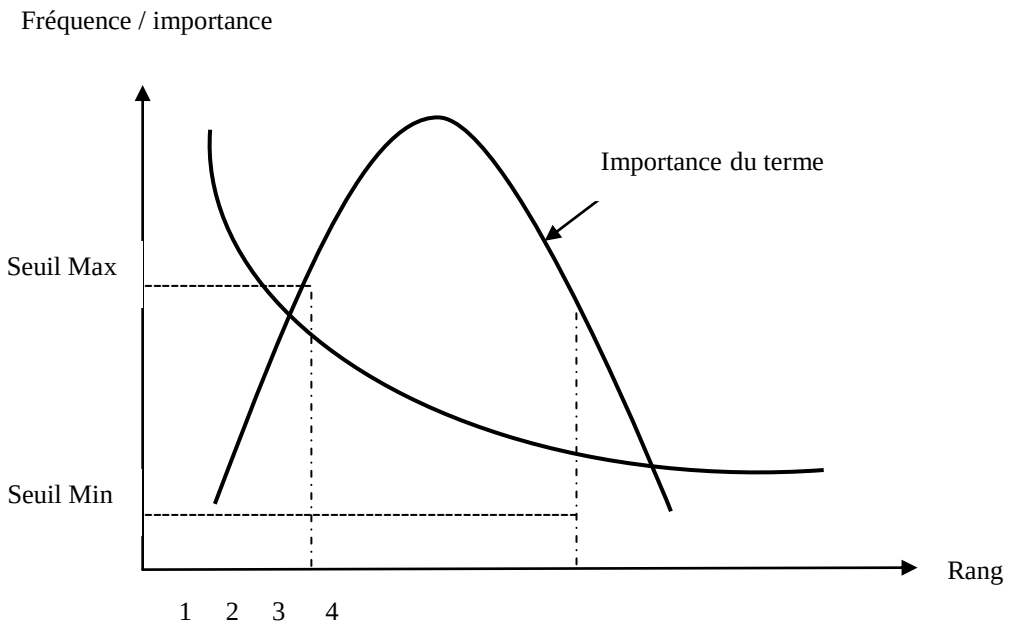

Figure I.3: Le rapport entre le rang × fréquence et l'importance du terme

III.2.4.2 La conjecture de Luhn

La conjecture de Luhn est une mesure basée sur la loi de Zipf. Donc l'importance d'un terme dans un document est lie à sa fréquence ou encore sont pouvoir expressif (appeler aussi l'informativité). Elle considère que les termes qui ont un rang faible ou élevé sont les termes non pertinents (informativité limite), et les termes qui ont un rang moyen sont considérés comme des termes pertinents (meilleure informativité). Cette conjecture est utilisée pour diminuer la taille des index des documents.

La correspondance entre le pouvoir expressif et la fréquence d'apparition dans un document est illustrée dans la figure suivante [Baziz, 2005] :

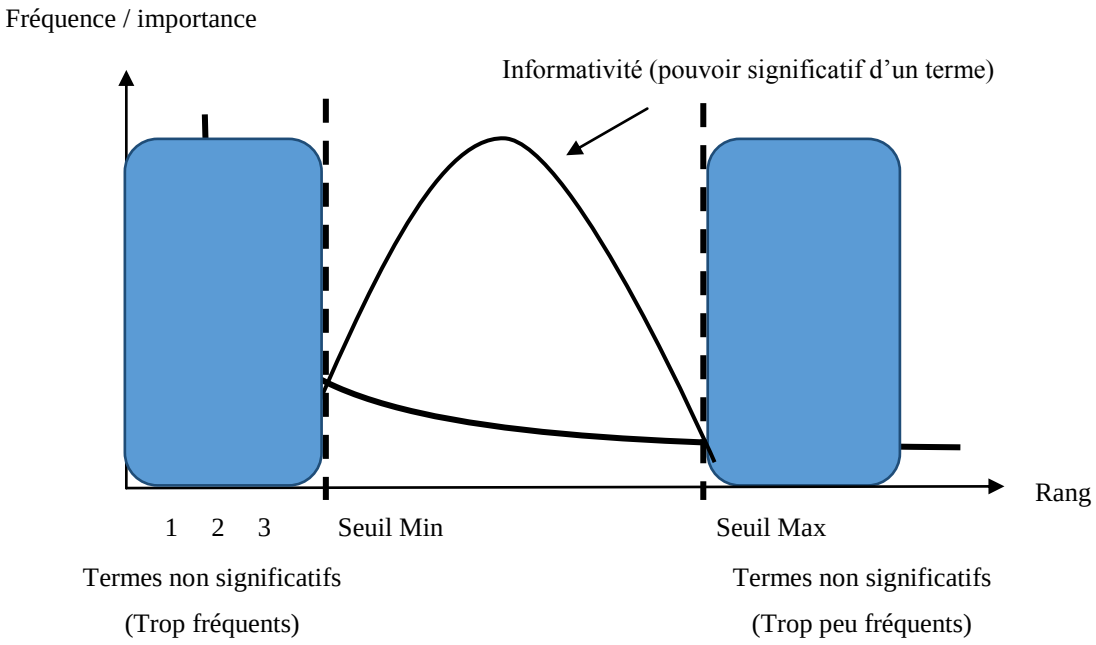

Figure I. 4: La conjecture de Luhn

III.2.4.3 Le TF (Term Frequency)

La fréquence du terme est une pondération locale, elle est simplement une fonction de fréquence de ce terme t_i dans le document d_j c'est-à-dire le nombre d'occurrences d'un terme dans le document. Cette pondération est notée souvent $\boldsymbol{tf_{ij}}$.

Elle peut être utilisée selon plusieurs déclinaisons

- **Fonction brute :**

  TF

- **Fonction binaire** *:*

  [0/1] : présence, absence

- **Fonction logarithmique :**

  $a + log(Tf)$ *Où α est une constante.*

- **Fonction normalisée** :

  $\frac{Tf}{Max\ (Tf)}$ Ou $0.5+0.5\frac{Tf}{Max\ (Tf)}$

Où max (Tf) est la plus grande valeur tf des termes du document D_j [Baziz, 2005]

III.2.4.4 IDF (Inverse document Frequency)

La fréquence inverse de document est une pondération globale, mesure l'importance d'un terme dans toute la collection. Elle prend en considération le nombre total des documents de la collection (corpus) et le nombre de documents contenant ce terme [these]. Les termes qui ont un meilleur poids ce sont les termes les moins fréquents dans la collection, c'est-à-dire les termes qui sont présent dans des nombreux documents sont les moins utiles [Baziz, 2005].

Cette mesure (IDF) est utilisée selon l'une des déclinaisons suivantes :

$$\text{Log}\left(\frac{N}{n_i}\right) \qquad \text{Ou} \qquad \text{Log}\left(\frac{N-n_i}{N}\right)$$

Où

n_i : le nombre de documents contenant le terme t_i

N : le nombre total de documents dans la collection.

III.2.4.5 Le TF-IDF (Inverse document Frequency)

La mesure TF-IDF consiste à multiplie les deux mesures TF×IDF, elle donne une bonne représentation du poids, donc une bonne approximation de l'importance du terme dans le document. Ainsi, un terme qui a une valeur de tf*idf élevée doit être à la fois important dans ce document, et aussi il doit apparaître peu dans les autres documents. C'est le cas où un terme correspond à une caractéristique importante et unique d'un document [Nassr, 2002].

Cette mesure est donnée comme suit :

$$TF \times IDF = log\ (1 + TF) \times IDF$$

III.2.5 Le résultat de l'indexation : L'index

Apres l'analyse lexicale, l'élimination des mots vides, la normalisation, la pondération. Le résultat de ces phases (l'index) est l'ensemble des descripteurs et leurs pondérations qui vont représenter au mieux le contenu de ces document (ou bien requête), ces termes peuvent être soit un mot, soit une racine de mot, soit un terme composé, etc….

Ce résultat peut représenter comme suit :

$$D\text{j} \rightarrow \{\ldots, (t_{\text{i}}, a_{\text{ij}}), \ldots\}$$

Où

- ***t*** le terme d'indice **i** dans le vocabulaire
- $\boldsymbol{a_{ij}}$ son poids dans le document $\boldsymbol{D}_{\text{j}}$.

Avec cette structure, il est facile de trouver les termes inclus dans un document. [Yaël, 2009]

III.3 Exemple d'indexation

- **Document original (D) :**

"The present study is a history of the DEWEY Decimal Classification. The first edition of the DDC was published in 1876, the eighteenth edition in 1971, and future editions will continue to appear as needed. "[Yaël, 2009]

- **Après analyse lexicale :**

The present study is a history of the Dewey decimal classification the first edition of the ddc was published in 1876 the eighteenth edition in 1971 and future editions will continue to appear as needed

- **Après suppression des mots vides :**

present study history dewey decimal classification edition ddc published 1876 eighteenth edition 1971 future editions continue needed

- **Après radicalisation avec l'algorithme Porter :**

present studi histori dewey decim classif edit ddc publish 1876 eighteenth edit 1971 futur edit continu need

- **Le résultat de l'indexation : l'index :**

$$D\text{j} \rightarrow \{\ldots ; (t\text{i}, a\text{ij}) ; \ldots\}$$

Donc,

d1 → {(edit , 0.090) ; (dewey, 0.25); (decim, 0.125); (classif, 0.019); (present, 0.003); (studi, 0.002); (histori, 0.039); (publish, 0.008); (ddc, 0.4); (eighteenth, 1.0); (futur, 0.010); (continu, 0.014); (need, 0.022)}.

IV. Recherche ou L'appariement document / requête

Dans un SRI une fois les documents transformés (processus d'indexation), l'utilisateur peut interroger la collection de documents à travers la formulation d'une requête qui exprime son besoin informationnel, cette requête est représentée sous une forme interne compréhensible par le système (processus d'indexation) [Ressad-Bouidghaen, 2011].

Le processus d'appariement est le résultat de la comparaison entre les documents et la requête qui donne une liste de documents, ces documents sont ordonnés selon un score de similarité qui est donné par une fonction nommée Retrieval Status Value. Elle est notée RSV(d,q), où d est un document et q est une requête [Yaël, 2009].

Ce score s'appuie sur des approches mathématiques. On en distingue :

– le modèle booléen ou ensembliste,

– le modèle vectoriel,

– le modèle probabiliste,

– les réseaux inferentiels bayésiens,

– le modèle connexionniste,

– les modèles de langage,

– latent Semantic Indexing : LSI,

Dans ce qui suit, nous détaillons quelque modèle.

V. Mécanismes de reformulation de requête

La qualité d'un SRI dépend de sa capacité à retrouver des documents pertinents pour l'utilisateur, pour garantir cet objectif on reformule généralement les requêtes. La reformulation de requête consiste à créer une nouvelle requête plus adéquate que celle initialement formulée par l'utilisateur, et par conséquent le SRI augmentera son rappel et sa précision [Ressad-Bouidghaen, 2011].

Concrètement la reformulation de la requête consiste à ajouter, supprimer et/ou pondérer les termes (enrichir et affiner la requête utilisateur), reliés à ceux de la requête initiale. [Yaël, 2009]

La reformulation elle est donc liée au choix des termes par le système (index) et par l'utilisateur (requête). On peut distinguer trois types de reformulation [Ressad-Bouidghaen, 2011] :

V.1 La reformulation manuelle

Ce type reformulation est généralement utilisé dans les systèmes de recherche booléens. Donc pour reformuler la requête initiale pour faire une nouvelle recherche des documents pertinents on a besoin d'utilisant un vocabulaire contrôlé (thésaurus ou classification), pour trouver les bons termes pour compléter la requête [Ressad-Bouidghaen, 2011].

V.2 La reformulation semi-automatique (interactive)

C'est la reformulation la plus populaire qui appelée aussi réinjection de la pertinence (**relevance feedback**). Le principe de cette stratégie est de présenté une liste de documents à l'utilisateur, ils sont sélectionnés par le système (les documents jugé pertinent) comme réponse à la requête de l'utilisateur (la requête initiale). Après l'utilisateur c'est lui qui va examiner ces documents et décide du choix des termes à ajouter dans la requête (indiquer ceux qui sont pertinents à lui) [Ressad-Bouidghaen, 2011].

V.3 La reformulation automatique

Appelé aussi le pseudo réinjection de la pertinence, dans ce cas la reformulation est faite d'une manière automatique sans l'intervention de l'utilisateur, soit par l'utilisation d'une ressource externe qui peut être un thesaurus qui regroupe plusieurs informations de type linguistique (équivalence, association, hiérarchie) et statistique (pondération des termes), une ontologie, etc…, soit par l'exploitation des n premiers documents renvoyés par le système comme pertinent à la réponse de la requête initiale (blind feedback) [Ressad-Bouidghaen, 2011].

« Le problème avec la reformulation automatique est l'estimation des « bons » termes qui peuvent conduire effectivement à une amélioration du processus de recherche car l'introduction des termes inappropriés peut entraîner un silence ou au contraire augmenter un bruit. » [Aliane et al, 2007]

VI. Quelques modèles de la recherche d'information

Dans [Daoud, 2009] « Un modèle de RI a pour rôle de fournir une formalisation du processus de RI et un cadre théorique pour la modélisation de la mesure de pertinence. »

Étant donné un ensemble de termes pondérés issus de l'indexation, le modèle remplit deux rôles suivants :

- Créer une représentation pour un document ou pour une requête basée sur ces termes. [Yaël, 2009]
- définir une méthode pour comparer la représentation de document et la représentation de requête afin de déterminer leur degré de correspondance (ou similarité). [Yaël, 2009]
- Les modèles courants peuvent être classés en trois modèles principaux [Baziz, 2005]
- Les modèles basés sur la théorie des ensembles : (modèle booléen et le modèle booléen étendu).
- Les modèles algébriques : (modèle vectoriel).
- les modèles probabilistes.

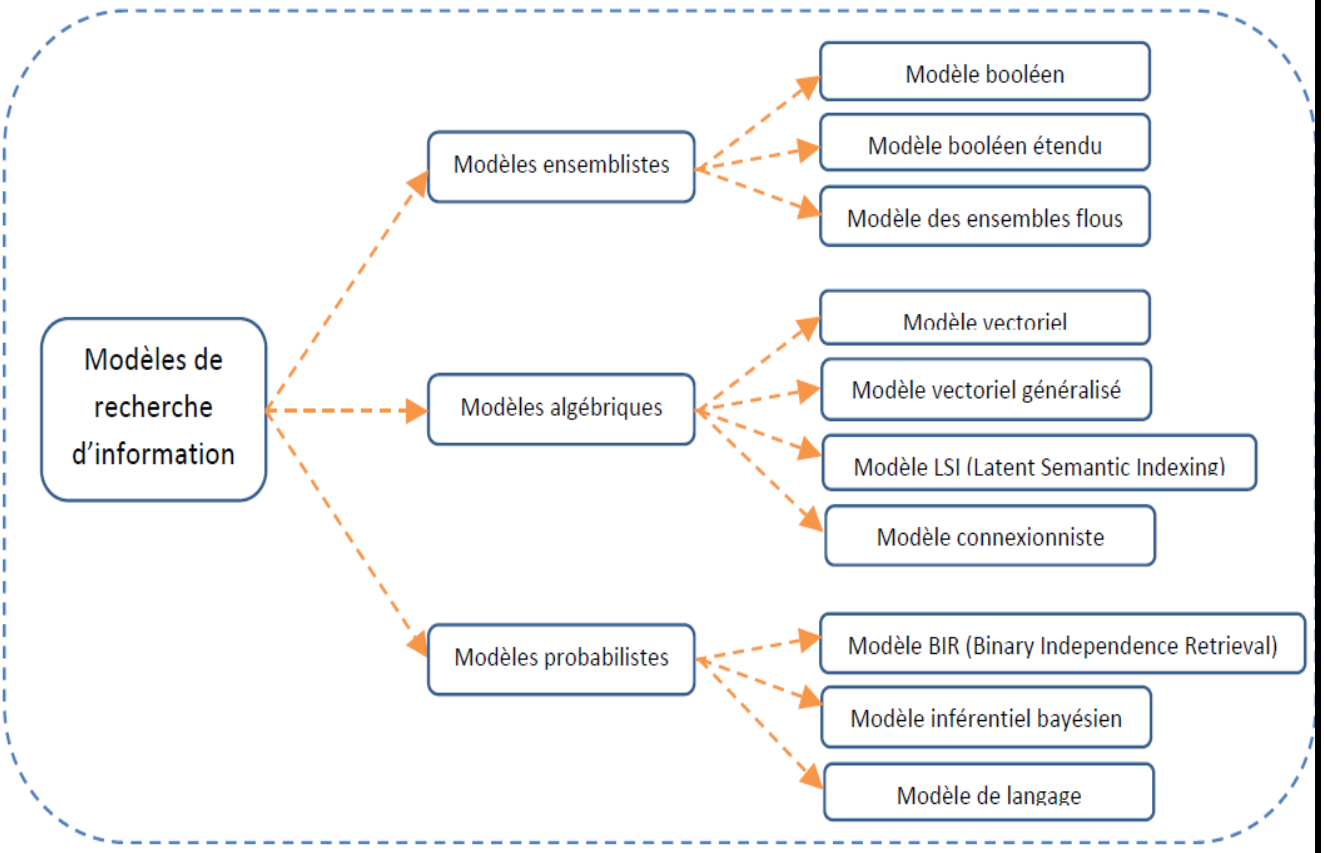

Figure I.5 : Les modèles de RI [Chebili, 2011]

VI.1 Les modèles ensemblistes

VI.1.1 Le modèle booléen

C'est le premier modèle utilisé dans la recherche d'information [Nassr, 2002] .Dans ce modèle les documents sont représentés par une conjonction logique des termes, un document peut être s'écrit comme suit : d = t1 ∧ t2 ∧ t3... ∧ tn.

La requête est représentée sous forme d'expression logique des termes [Ho, 2004]. Les termes sont reliés par des opérateurs logiques ET(∧), OU(∨) et NON(¬)

Ex : $Q = (q_1 \wedge q_2) \vee (q_3 \wedge \neg q_4)$.

La similarité RSV (D, Q) entre une requête et un document est définie de la façon suivante :

RSV(D_j,q_i)=1 si $q_i \in D_j$; 0 sinon,

RSV(D_j,$q_i \wedge q_j$)=1 si RSV(D_j,q_i)=1 et RSV(D_j,q_j)=1 ; 0 sinon,

RSV(D_j,$q_i \vee q_j$)=1 si RSV(D_j,q_i)=1 ou RSV(D_j,q_j)=1 ; 0 sinon,

RSV(D_j, $\neg q_i$)=1 si RSV(D_j,q_i)=0 ; 0 sinon,

Les inconvénients de ce modèle :

- La présence de deux valeurs (1 ou 0) pour calculer la similarité entre le document et la requête, une liste non-ordonnée retournée par le système, difficile de dire qu'un document est mieux qu'un autre.
- Les termes sont pondérés de la même façon 1 ou 0 difficile de juger qu'un terme important que l'autre c.-à-d. les termes ont la même importance.
- Les utilisateurs trouvent une difficulté d'exprimer le besoin avec les opérateurs logiques c'est-à-dire manipulent très mal les opérateurs booléen.

VI.1.2 Le modèle booléen étendu

Ce modèle est une extension du modèle booléen est introduit par Salton en 1983 [Benaouicha, 2009] l'idée est de donner un poids chaque terme du document, « Il tient compte de l'importance des termes dans la représentation des documents » [Baziz, 2005] tout en proposant une pertinence graduée, dans [Benaouicha, 2009] « Ce modèle peut être vu comme une combinaison des modèles booléen et vectoriel ». Le poids d'un terme dans un document est une valeur comprise entre 0 et 1.la requête reste toujours une expression logique.

Considérons un ensemble de termes $\{t_1, \ldots, t_N\}$ et soit wd_{ij} le poids du terme t_i dans le document d_j, où $D_j = (wd_{1j}, \ldots, wd_{Nj})$, avec $1 \leq i \leq N$ et $0 \leq wd_{ij} \leq 1$ [Chebili, 2011]

La correspondance entre une requête et un document est définie de la façon suivante :

Opérateur OU : $$RSV(D_j,Q_k) = \left(\frac{\sum_{i=1}^{n} (wq_{ij}{}^{p} \times wd_{ij}{}^{p})^{\frac{1}{p}}}{\sum_{i=1}^{n} wq_{ik}{}^{p}} \right)^{\frac{1}{p}}$$

Opérateur ET : $$RSV(D_j,Q_k) = \left(\frac{\sum_{i=1}^{n} (wq_{ij}{}^{p} \times (1 - wd_{ij}{}^{p}))^{\frac{1}{p}}}{\sum_{i=1}^{n} wq_{ik}{}^{p}} \right)^{\frac{1}{p}}$$

Où : $0 \leq p \leq 1$ est une constante, et $wq_{ik}{}^{p}$ le poids du terme t_i dans la requête Q_k.

VI.2 Les modèles algébriques

VI.2.1 Modèle vectoriel

Dans ce modèle les documents et les requêtes sont représentés par des vecteurs dans un espace vectoriel de dimension N engendré par les termes d'indexation [Baziz, 2005].

Soit l'espace vectoriel défini par l'ensemble des termes : $<t_1, t_2, \ldots, t_n>$.

Un document et une requête sont représentés par les vecteurs suivants :

$$D_j=< d_{1j}, d_{2j}, \ldots, d_{nj} > \qquad Q=< q_1, q_2, \ldots, q_n >$$

Où d_{kj} et q_k représente le poids de terme t_k dans le document D_j et la requête Q avec k=1..n

Les composants la requête Q sont pondérés de la même façon que celle du document D_j

Les principales mesures de similarité utilisée sont [Baziz, 2005] :

Le produit scalaire :

$$RSV(D_j, Q) = \sum_{i=1}^{N} q_i \times d_{ij}$$

Mesure de Jaccard :

$$RSV(D_j, Q) = \frac{\sum_{i=1}^{N} q_i \times d_{ij}}{\sum_{i=1}^{N} q_i{}^2 + \sum_{i=1}^{N} d_{ij}{}^2 - \sum_{i=1}^{N} q_i \times d_{ij}}$$

La mesure cosinus :

$$RSV(D_j, Q) = \frac{\sum_{i=1}^{N} q_i \times d_{ij}}{\sqrt{\sum_{i=1}^{N} q_i{}^2} \times \sqrt{\sum_{i=1}^{N} d_{ij}{}^2}}$$

VI.3 Les modeles probabilistes

Ce modèle est basé sur le principe de classement en utilisent les probabilités, un SRI basé sur ce modèle a pour but classer les résultats de la recherche en fonction de leur

probabilité de pertinence de document par rapport à la requête . Il y a deux types de documents : les pertinents (R) et les non pertinents (NR)

La pertinence entre le document D_j et la requête Q est déterminée par :

$$RSV(D_j, Q) = \sum_{i=1}^{t} log \frac{P(w_{ij}/R)}{P(w_{ij}/NR)}$$

- $P(w_{ij}/Pert)$: probabilité que le terme t_i de poids w_{ij} occurre dans le document D_j sachant que ce dernier est pertinent pour la requête.
- $P(w_{ij}/NonPert)$: probabilité que le terme t_i occurre dans le document Dj sachant que ce dernier n' est pas pertinent pour la requête.

VII. Evaluation d'un système

L'évaluation de performance des **SRI** (modèles et méthodes) a toujours été un centre d'intérêt très important dans le domaine de la Ri, tout cela pour atteindre un SRI idéal qui ramène tous les documents pertinents de collection qui répond au besoin en information de l'utilisateur, et rejette les documents non pertinents. La qualité des **SRI** est relie avec les réponses du système qu'a été souhaitée par l'utilisateur. Plus les réponses du système correspondent à celles que l'utilisateur espère, meilleur est le système [Karbasi, 2007].

Dans un SRI, l'objectif est de minimiser le bruit (les documents non pertinents restitués), et le silence (les documents pertinents non restitués).

Donc l'évaluation des SRI est une étape importante, elle permet d'évaluer la performance du système. D'une part elle est utilisée pour paramétrer le SRI, d'estimer l'impact de chacune de ses caractéristiques et d'autre part elle est la base de la comparaison entre différents modèles de RI [Hlaoua, 2007].

« L'évaluation des systèmes peut être abordée selon deux angles : l'efficience et l'efficacité. » [Baziz, 2005]

➢ L'efficience regroupe le temps et l'espace :

Un système est considéré meilleur lorsque le temps entre la formulation de la requête et la réponse du système est court et l'espace occupé par le système est faible.

➢ L'efficacité d'un système peut être mesurée par les critères suivants :

- L'effort, intellectuel ou physique, nécessaire aux utilisateurs pour formuler les requêtes, conduire leur recherche, voir les documents résultats
- La présentation du résultat (capacité de l'utilisateur à utiliser les documents retrouvés)
- La qualité du corpus vis à vis du besoin de l'utilisateur (dans quelle mesure tous les documents pertinents sont dans le corpus)
- La capacité du système à retrouver des documents intéressants et à éliminer les autres. Cette caractéristique semble être la plus importante. [Baziz, 2005]

VIII. Corpus de test

Le corpus ou la collection de test constitue le contexte d'évaluation. Pour réaliser une telle évaluation, on doit d'abord connaître les réponses idéales de l'utilisateur. Ainsi une expérimentation qui utilise les éléments suivants doit être établie (corpus de test):

- ✓ Un ensemble de documents.
- ✓ Un ensemble de requêtes.
- ✓ La liste de documents pertinents pour chaque requête.
- ✓ Des mesures et des critères quantifiables.

Les corpus diffèrent selon le domaine de spécialité, le nombre de documents de la collection, le nombre de requetés, la façon de juger la pertinence, etc…. Cette dernière est considérée comme la tâche la plus difficile.

Pour qu'un corpus de test soit significatif, il faut qu'il possède un nombre de documents assez élevé [Ressad-Bouidghaen, 2011].

De nombreuses et différentes collections (corpus) de test sont utilisées comme moyen pour l'évaluation des systèmes de recherches d'information. Les collections de test sont le résultat de projets d'évaluation qui se sont multipliés depuis les années 1970, on peut citer la collection CACM, la collection CISI, la compagne CLEF (Cross-Language Evaluation Forum), la campagne NTCIR, La campagne FIRE (Forum for Information Retrieval Evaluation), Le paradigme de Cranfield. Ou encore la campagne INEX. La campagne la plus connue est sans conteste TREC (Text REtrieval Conference) organisée annuellement depuis1992 par la NISI et la DARPA [Hammache, 2013].

Les campagnes d'évaluation sont un élément incontournable de la RI, elles fournissent des outils d'évaluation des systèmes, elles permettent la comparaison de systèmes et elles définissent des cadres d'études utiles aux chercheurs. [Yaël, 2009]

IX. Mesures d'évaluation

Dans un système de recherche d'information, le principal objectif est l'évaluation de la capacité du système à retrouver les documents pertinents pour une requête (minimiser le bruit et le silence). Cet objectif est évalué à l'aide de différentes mesures d'évaluation. Ces mesures sont souvent basées sur le rappel et la précision ou des mesures dérivées du rappel et de la précision. Nous citons :

- La précision à *n* documents restitués (noté : précision@n)
- La mesure F
- La précision exacte
- La précision interpolée
- La précision moyenne(ou MAP pou Mean Average Precision).

IX.1 Précision

C'est la capacité d'un système à sélectionner que des documents pertinents à une requête. Si le rappel vaut 1 c'est que les documents pertinents disponibles ont tous été sélectionnés par le système, inversement si le rappel vaut 0 c'est qu'aucun document pertinent n'a été sélectionné. Cette mesure permet aussi de déterminer le silence, c'est-à-dire les documents pertinents non retrouvés. [Yaël, 2009]

IX.2 Rappel

C'est la capacité d'un système à sélectionner tous les documents pertinents de la collection pour une requête. La précision vaut 1 quand tous les documents sélectionnés sont pertinents. Elle vaut 0 si aucun des documents sélectionné n'est pertinent. Cette mesure détermine également le bruit, c'est-à-dire les documents non pertinents retrouvés par le système de recherche. [Yaël, 2009]

Ces deux mesures sont données par les formules suivantes :

$$Précision = \frac{|DRP|}{|DR|} \qquad Rappel = \frac{|DRP|}{|DR|}$$

Avec :

- DRP : l'ensemble des documents pertinents vis-à-vis de la requête et qui sont retournés par le SRI,
- DR : l'ensemble des documents retournés par le SRI,
- DP : l'ensemble des documents de la collection qui sont pertinents vis-à-vis à la requête,
- |DRP|, |DR|, |DP| : la cardinalité des ensembles considérés (le nombre de documents).

Le Silence = 1 - Rappel et le Bruit = 1 - Précision. [Yaël, 2009]

La Figure I.7 [Harrathi, 2010] illustre la répartition des documents suite à une interrogation utilisateur.

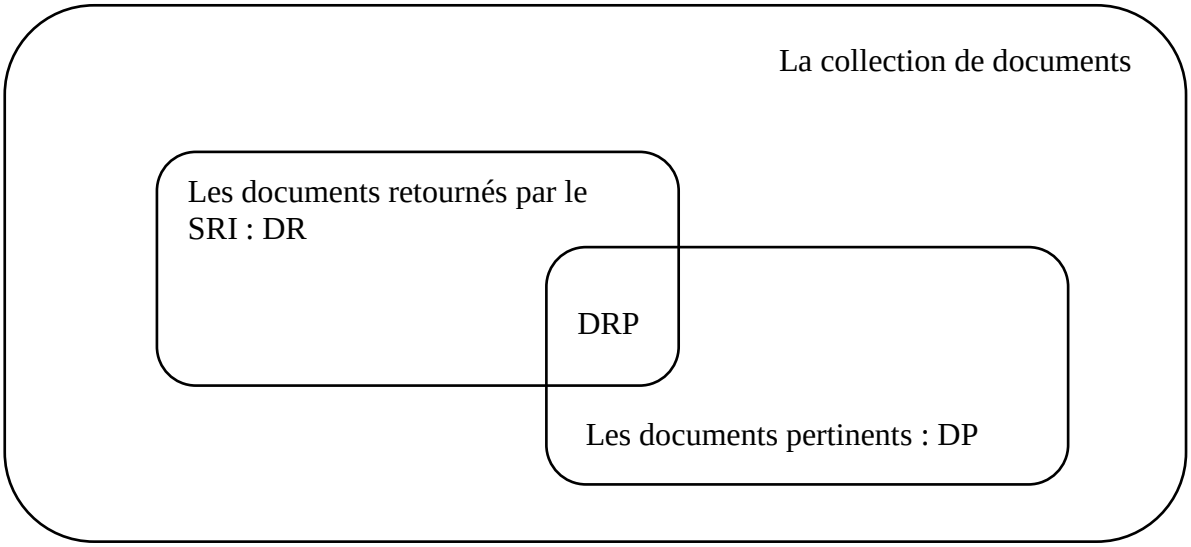

Figure I.6: La répartition des documents par rapport au besoin d'utilisateur

Il y a une forte relation entre le rappel et la précision : quand l'une augmente, l'autre diminue. Il ne signifie rien de parler de la qualité d'un système en utilisant seulement une des mesures. Ainsi, pour un système, on a une courbe de précision-rappel qui a en général la forme suivante :

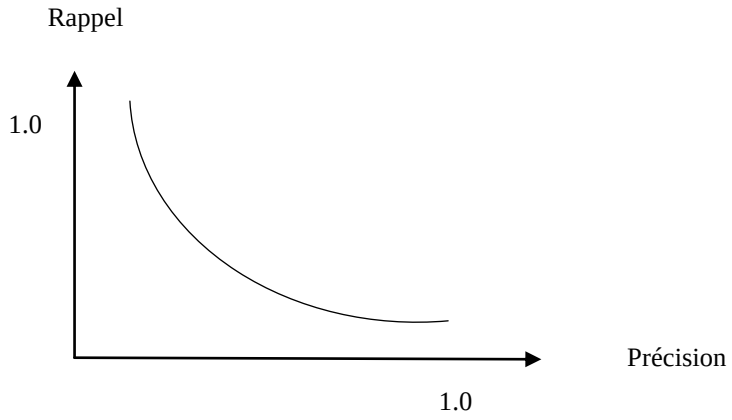

Figure I.7: Courbe Rappel/Précision

IX.3 La précision moyenne

La précision moyenne est une mesure de performance globale. La précision moyenne est une moyenne de précision sur un ensemble de points de rappel.

$$MAP = \frac{1}{n}\sum_{i=1}^{N} P(i) * R(i)$$

Avec

- *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.
- *n* le nombre de documents pertinents restitués.
- N le nombre total de documents. [Yaël, 2009]

IX.4 Précision à n (P_n)

C'est la précision à différents niveaux de coupe. Cette précision mesure la proportion des documents pertinents retrouvés parmi les n premiers documents retournés par le système. Cette mesure ne prend pas en compte le que certaines requête comportent peu de documents pertinents. [Daoud, 2009]

$$P_n = \frac{\#Documents\ Pertinents\ Retrouvés\ (DPR)}{n}$$

IX.5 La précision exacte

Notée aussi R-précision, cette précision mesure la proportion des documents pertinents retrouvés après que R documents ont été retrouvés, ou R est le nombre de documents pertinents pour la requête considérée. C'est une variante de la $\mathbf{P_n}$ où n est le nombre de documents pertinents pour la requête [Hammache, 2013].

X. SRI et Web social

L'évolution d'Internet a permis la création de nouveaux outils de communication et de recherche. Ces outils ce sont les réseaux sociaux (RS).qui sont maintenant reconnus comme un moyen important pour la diffusion de l'information. L'explosion des RS a permis l'émergence d'une nouvelle branche de la Recherche d'Information (RI) : la RI social.

Les utilisateurs du Web produisent divers contenus, créent des annotations, manipulent les documents sur le Web et laissent des traces de leurs passages, etc. ces données (générées par les utilisateurs) peuvent très utiles pour faire la recherche et la gestion de l'information sur le web. Elles permettent par exemple de booster les performances des moteurs de recherche actuels, mais elles nécessitent un processus de préparation et de nettoyage qui peut être fastidieux. Cette étape est similaire au processus de préparation des entrepôts de données.

Nous notons l'existence de deux réseaux Flickr et Delicious qui s'intéressent à l'intégration des annotations dans les SRI (le but est de faciliter l'accès à l'information)

- **Flickr**

Flickr[2] Est un site web social permet stocker, partager des photos, de les rendre publiques, là où n'importe quel utilisateur inscrit peut stocker jusqu'à 1000 Go des photos gratuit, il peut aussi ajouter des commentaires, des annotations, créer une collection d'images, participer à des groupes. Flickr dispose d'un outil de recherche optimisé permet de trouver les photos à tout moment

[2] https://www.flickr.com/

- **Delicious**

Delicious[3] est un site web qui permet de sauvegarder ses marque-pages, de les classer selon le principe de folksonomie par mots clés (aussi appelé tag) et aussi et surtout de pouvoir les partager avec d'autres utilisateurs, aussi permet d'accéder à ces favoris de n'importe quelle machine via le compte d'utilisateur. Cela permet entre autre de faire des recherches dans les favoris identifiés par les utilisateurs. Delicious propose un moteur de recherche très performant sur l'ensemble des liens publiques bookmarkés par ses utilisateurs.

En tapant un tag dans la barre de recherche s'affiche en dessous de cette barre tous les tags les plus utilisés similaires à la requête. Cela est très pratique pour deux raisons. La première concerne l'orthographe d'un mot, cela permet de ne pas se tromper notamment lorsqu'il s'agit de noms propres. La deuxième permet de se rendre compte des termes utilisés par la communauté des utilisateurs de Delicious.

XI. Conclusion

Nous avons présenté dans ce chapitre les principales notions et concepts de base de la recherche d'information. Nous avons développé les principales étapes d'un processus de recherche d'information que sont la représentation ou indexation de l'information et la comparaison de l'information et du besoin en information, ainsi que les principaux modèles de recherche d'information, de plus la technique d'évaluation d'un système de recherche d'information , et pour terminer la relation entre SRI et un Web social.

[3] https://delicious.com/

Chapitre II :
Les générations du Web

I. Introduction

Le web est sans nul doute une technologie majeure du 21ème siècle. Et si sa nature, sa structure et son utilisation ont évolué au cours du temps, force est de constater que cette évolution a également profondément modifié nos pratiques commerciales et sociales.

« Le Web n'est plus une collection de pages statiques en HTML qui décrivent quelque chose du monde. De plus en plus, le Web est le monde – chaque chose et chaque personne de ce monde projettent une « ombre d'information », une aura de données, qui, captée et traitée de manière intelligente, ouvre d'extraordinaires possibilités et de stupéfiantes implications. Le Web puissance deux est notre façon d'explorer ce phénomène et de lui donner un nom » [Audet, 2010].

Depuis sa création en 1989 par Tim berners-Lee (considéré aujourd'hui comme le père fondateur du Web), le « Web » ((nom anglais signifiant «toile d'Araignée Mondiale»), contraction de «World Wide Web» (d'où l'acronyme www)), est devenu très rapidement à la fois le service le plus populaire d'Internet et la plus grande base de données existante. Son contenu a très rapidement évolué et ce à plusieurs niveaux :

- La quantité d'information (plus de 30 milliards pages web)
- Le nombre d'utilisateurs (plus de 2,3 milliards d'internautes dans le monde)
- La nature et la structure du web (du Web 1.0 au Web 2.0 puis au Web 3.0, 4.0 etc...) [Pruski, 2009].

Le Web est une des applications d'Internet, comme le sont le courrier électronique, les réseaux sociaux, la messagerie instantanée et les systèmes de partage de fichiers poste à poste.

II. Web 0.0

Le Web 0.0 ou **Web militaire** symbolise l'origine de l'Internet avec la création en 1972 du premier réseau de données à transfert de paquets (ARPANET)[4] suite à la commande du Pentagone portant sur la création d'un réseau capable de résister à une attaque militaire [Trudeau, 2010].

[4] http://hautrive.free.fr/reseaux/architectures/reseaux-arpanet.html

III. Web 1.0

Le **web 1.0** celui des années 90, encore appelé **web traditionnel**, est avant tout un web statique. Il représente les sites de première génération. Les contenus (texte/image/vidéo/son) sont produits et hébergés par une entreprise, propriétaire du site. Le web traditionnel est centré sur la distribution d'informations, C'est un web passif : l'internaute y consomme de l'information, comme on peut le faire dans une bibliothèque par exemple (L'utilisateur n'est que lecteur de l'information). Ce web comprenait des pages statiques reliées entre elles par des liens hypertextes rarement mises à jour [Chaimbault, 2007].

Formellement le web de cette version est vu comme un graphe dont les nœuds sont des pages (généralement statiques) et les arêtes représentent les liens hypertextuels.

Au milieu des années '90, Avec l'apparition de nouveaux langages de scripts et du DHTML (comme le PHP ou l'ASP) couplés avec une base données, certains sites deviennent alors dynamiques (le web 1.5 ou les« dot-com»). C'est-à-dire que le contenu (texte, image, vidéo, son) est géré par un système de gestion de contenu ou CMS (Content Management System). Ils permettent alors à plusieurs individus de travailler et de modifier les informations sur un même document. Ces technologies autorisent aussi la séparation de gestion de la forme et du contenu [Chaimbault, 2007].

Dans le web 1.5 les technologies ont un peu changé par rapport au web 1.0, le web se voit moins statique, mais la logique fondamentale centrée sur l'importance des produits du web proposés aux usagers demeurait la même.

- La notion de « site Web » (l'individu ne peut pas modifier une information mais uniquement la consulter : l'exemple d'une bibliothèque),
- Le propriétaire du site est le seul qui a la possibilité de publier du contenu sur l'Internet,
- des pages dynamiques en PHP et des pages statiques en HTML,
- l'attitude passive de l'internaute qui ne peut que consulter les pages [Trudeau, 2010].

IV. Web 2.0

IV.1 Introduction

“The new Web is a very different thing. It's a tool for bringing together the small contributions of millions of people and making them matter. Silicon Valley consultants call it Web 2.0, as if it were a new version of some old software. But it's really a revolution. « Time's Person of the Year: You”», [O’Reilly & al, 2009]

Quand le web 1.0 reproduit un modèle de communication dit "one to many" commun aux médias traditionnels (télévision, radio, presse), les dispositifs socio-techniques 2.0 proposent de nouveaux usages reposant sur un modèle de communication "many to many″ [Quoniam & al, 2009].

Alors, l’internet et ses utilisateurs ont subi un important changement qui a déterminé le passage vers un nouveau modèle : qu’on appelle Web 2.0 ou Web participatif ou web collaboratif ou encore **web social,** qui est une nouvelle vision d’internet visible partout dans le monde et dans lequel n'importe quel internaute peut être actif. Il a permis de transformer des internautes en écrivains du Web.

Le rêve de Tim Berners-Lee devient réalité : les internautes ne sont plus seulement consommateurs passifs, mais contribuent activement d'une part à la création de contenus, mais aussi à la validation de leur valeur.

L’expression « web 2.0 » est inventé par D. Dougherti en aout 2001 de la société O’Reilly Média et rapidement rendu populaire par Tim O’Reilly en octobre 2004 lors de la première conférence sur le web 2.0, pour désigner une mutation du web statique vers le web participatif. [Crepel, 2011] Ce phénomène (le web 2.0) marque l'apparition d'un nouveau paradigme de communication, qui est former la façon dont nous travaillons et interagissons avec l'information sur internet. « Plus qu’un réseau reliant des pages, le Web d’aujourd’hui relie des gens ; d’où l’appellation web social » *Martin Lessard*.

IV.2 Définition

La définition ci-dessous prise de Wikipédia décrit cette évolution au web 2.0 :

« Le Web 2.0 est l'évolution du Web vers plus de simplicité (ne nécessitant pas de connaissances techniques ni informatiques pour les utilisateurs) et d'interactivité (permettant à chacun, de façon individuelle ou collective, de contribuer, d'échanger et de collaborer sous différentes formes). L'expression « Web 2.0 » désigne l'ensemble des

techniques, des fonctionnalités et des usages du World Wide Web (www) qui ont suivi la forme originelle du web, en particulier les interfaces permettant aux internautes ayant peu de connaissances techniques de s'approprier les nouvelles fonctionnalités du web. Ainsi, les internautes contribuent à l'échange d'informations et peuvent interagir (partager, échanger, etc.) de façon simple, à la fois avec le contenu et la structure des pages, mais aussi entre eux, créant ainsi notamment le Web social. L'internaute devient, grâce aux outils mis à sa disposition, une personne active sur la toile ».

IV.3 Les 7 principes du web 2.0 [Chaimbault, 2007]

- Le web vu comme une plate-forme de services.
- Tirer parti de l'intelligence collective.
- La puissance est dans les données.
- La fin des cycles des releases.
- Des modèles de programmation légers.
- Le logiciel se libère du PC.
- Enrichir les interfaces utilisateurs.

IV.4 Glossaire du Web 2.0

De nombreux outils sociaux [Asselin, 2008] sont disponibles comme les wikis, les blogs, les systèmes de messages instantanés et de visio-conférences. Ces outils sociaux sont utilisés par des larges communautés produisant une grande quantité d'informations. Pour mieux comprendre les concepts utilisés dans l'univers du Web 2.0, nous vous proposons ce glossaire des termes essentiels :

IV.4.1 Tags (nuage de tags)

Étiquette, mot-clé ou graffiti Marqueur sémantique, qui permet de qualifier un contenu. Les tags se présentent souvent en « nuage », c'est-à-dire dans une liste non linéaire qui met en évidence la « popularité » d'un tag [Crepel, 2011].

IV.4.2 Folksonomie

« Une folksonomie F est un tuple F = (U, T, D, A) où U est l'ensemble des utilisateurs, T est l'ensemble des étiquettes, D est l'ensemble des documents web, et A ⊆ U ×T ×D est l'ensemble des annotations. » [Rupert, 2009].

Inventé par Thomas Vander Wal, le terme de folksonomy provient de la contraction des mots folks (« les gens ») et taxonomy (« taxinomie » ou « taxonomie » pour évoquer la notion de classification). La folksonomy décrit donc une pratique qui consiste à classer du contenu - de manière collaborative - à partir de tags (ou mots-clés) proposés par les internautes eux-mêmes [Crepel, 2011].

IV.4.3 Blog

Contraction de Web et Log, le blog est un journal en ligne qui permet à son animateur d'échanger ses points de vue avec ses lecteurs. En effet, chaque nouvel article peut faire l'objet de nombreux commentaires postés par les visiteurs du site.

- **Blogueur :** personne qui publie sur un blog.
- **Blogosphère :** contraction de blog et biosphère, désigne l'ensemble de la communauté qui anime des blogs.

IV.4.4 RSS

(Flux RSS ou Fils RSS): Les fils RSS (Really Simple Syndication) sont des flux de contenus gratuits en provenance de sites Internet. Ils permettent d'afficher les nouveaux contenus publiés sur un site sans avoir à le visiter.

IV.4.5 Blogroll ou blogoliste

Liste de liens vers d'autres blogs, présentés par l'auteur d'un blog. On peut syndiquer sur une même page des billets venant de ces blogs via les formats RSS.

IV.4.6 Wiki

Un wiki (site dynamique) est un outil de gestion de site web qui permet aux utilisateurs de publier et modifier facilement du contenu.

IV.4.7 Crowdsourcing

Le crowdsourcing consiste à utiliser la créativité, l'intelligence et le savoir-faire d'un grand nombre d'internautes, et ce, au moindre coût. La traduction littérale de crowdsourcing est « approvisionnement par la foule ».

IV.4.8 Streaming

Transfert de données multimédia en continu sur Internet, et qui permet donc la lecture du média en direct sans téléchargement.

IV.4.9 Widget

Le mot widget recouvre deux notions distinctes en relation avec les interfaces graphiques. Il peut être considéré comme étant la contraction des termes Windows (fenêtre) et gadget. C'est un petit module, paramétrable et personnalisable, qui permet d'embarquer de l'information en le transportant vers le point de destination de son choix.

IV.4.10 Réseaux sociaux

Les réseaux sociaux [Mathilde, 2009] sont apparus avec la création de web 2.0.Un réseau social est un site internet permettant à l'internaute de s'inscrire et créer un compte appelé le plus souvent « profil ». Le réseau est dit social parce qu'il permet d'échanger avec les autres membres inscrits sur le même réseau : des messages publics ou privés, des liens hypertextes, des vidéos, des photos, des jeux… L'ingrédient fondamental du réseau social reste cependant la possibilité d'ajouter des « amis », et de gérer ainsi une liste de contacts. Les réseaux sociaux touchent un public extrêmement large. Les étudiants et les adolescents les adultes, principalement sur les réseaux sociaux spécialisés dans l'entreprise et le monde du travail. Facebook est un des réseaux qui est avant tout une entreprise et qui gagne beaucoup d'argent de diverses manières. Cependant tous ces réseaux, certes fort pratiques, ne sont pas accessibles depuis certaines parties du monde où ils peuvent être censurés ou même bloqués. Il faut faire attention car les réseaux sociaux bien que bénéfiques peuvent être parfois très dangereux, principalement concernant la vie privée.

IV.4.11 Les Mashups

Mashup ou Mash-up (Le terme se traduit en français par remixage, ou mosaïque) , est une composition d'application ou de sites web (outils collaboratifs en ligne) qui utilise et croise et exploite le contenu de plusieurs autres sources (applications ou sites) pour fournir ou proposer un nouveau produit ou service. Il s'agit bien souvent de réutiliser des données existantes dans un contexte non prévu à l'origine de la création de ces données. Cette capacité de mixage repose sur l'ouverte des API (interface de programmation qui permet de recourir aux fonctions et contenus d'un site web à partir de commandes externes) [Seilles, 2012].

Les Mashups sont aujourd'hui très utilisés comme outils de conception pour créer de nouvelles applications avec peu ou pas de programmation. Sur le Web, les Mashups ont grandement amélioré la créativité des concepteurs, en leur permettant de combiner rapidement et simplement des informations provenant de diverses sources puis de les intégrer dans de nouvelles applications [Luong, 2012].

- Par exemple Trivop.com combine des cartes GoogleMaps et des avis sur les hôtels issus du réseau social TripAdvisor [Zammar, 2012].

IV.4.11.1 Les types de Mashup

Il y a trois types principaux de Mashup: le mashup de consommateur, le mashup de données et le mashup de business. [Binh, 2008]

- Le type de mashup de consommateur est le mashup combinant des éléments de données à partir de plusieurs ressources différentes, et il les représente sur une interface graphique unique. Les exemples typiques de ce type de mashup sont les applications de Google Maps.
- Le type de mashup de données est le mashup combinant des éléments de données à partir des ressources similaires. L'exemple typique de ce type est Yahoo Pipes.
- Le type mashup de business est la combinaison de deux types au-dessus. Il combine des données des ressources différentes ainsi que des ressources similaires. Puis, il ajoute des relations entre eux. Enfin, le résultat est utilisé par une application de business.

IV.4.11.2 Les caractéristiques des Mashups

Les Mashups reposent sur les caractéristiques principales [Luong, 2012] suivantes :

- Réutilisation et intégration des codes et des contenus.
- Le logiciel en tant que service (en anglais Software as a service - SaaS).
- Do it yourself (DIY) : n'importe qui peut être un auteur.

V. Web 3.0

V.1 Introduction

Le Web est généralement utilisé comme une base de données mondiale pour la recherche. Les moteurs de recherche d'aujourd'hui ne peut pas rechercher plus précis, peut-être la raison principale est que la structure et la taille de Web actuel ne permet pas de faire des recherches plus précises et efficaces. La deuxième raison ne peut pas être éliminée : Web contient maintenant un grand nombre de documents, et ce nombre à chaque un ou deux ans sera doublé [Cai & al, 2003].

Le Web sémantique Ou « le Web de données » le but principal de ce Web est de minimiser l'intervention manuelle des humains dans la réalisation de leurs objectifs. Et pour ce faire, les machines doivent comprendre la sémantique, la signification de l'information sur le Web. Il étend le réseau des hyperliens entre des pages Web classiques par un réseau de lien entre données structurées permettant ainsi aux agents logiciel d'accéder d'une façon plus intelligente aux différentes sources de données contenues sur le Web et, de cette manière, d'effectuer des tâches (recherche, apprentissage, etc.) plus précises pour les utilisateurs. Technologies du Web sémantique permettent aux gens de créer des banques de données sur le Web, de construire des vocabulaires, et écrire des règles pour le traitement des données. Ces données liées sont habilités par des technologies comme RDF, SPARQL, OWL, SKOS et.

V.2 Définition

Le « Web sémantique » est un terme inventé par Tim Berners-Lee qui en dit : « Le Web sémantique est une extension du Web actuel, dans lequel l'information a un sens bien défini, et permet une meilleure coopération dans le travail entre les humains et les

ordinateurs…Le Web des données qui peuvent être traitées par les ordinateurs »[W3C, 1998]

V.3 Architecture du web sémantique

Le schéma suivant [Seilles, 2012] représente l'architecture globale du web sémantique :

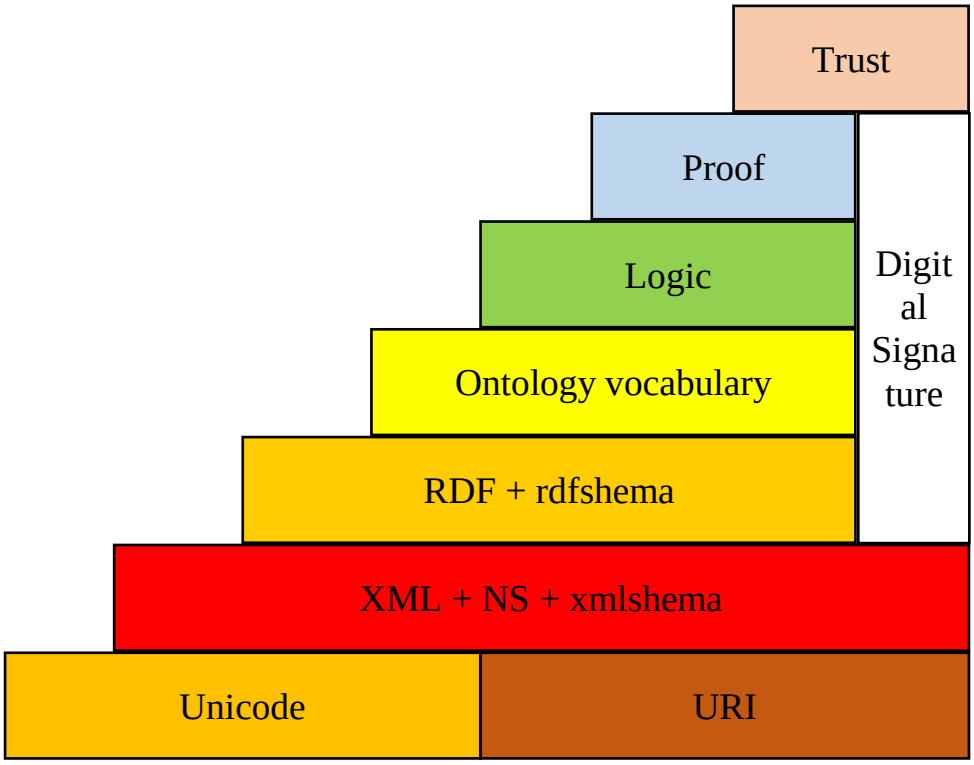

Figure II. 1 l'architecture du web sémantique

- **La couche URI/Unicode** : les données sont codées par le standard Unicode, ces données possèdent une adresse URI (Uniforme Ressource Identifier) (on distingue les url qui sont des id (ou adresse) dépendantes du réseau internet et les uri qui sont des id indépendants du réseau ISSN, ISBN.

- **XML/XML schéma** : Pour donner une bonne structuration à ces données, le Web Sémantique utilise les langages de balises normalisé comme XML (extensible Markup Language). Ainsi que des grammaires qui vérifient la correction syntaxique des annotations, métadonnées ou une ontologie

- **RDF/RDFs** : Resource Description Framework, Le but est de donner une organisation plus structurée des informations présentes sur le Web à travers une description sémantique des données fournie par XML. Pour réaliser ces objectifs RDF permet l'expression de métadonnées sur les ressources du web, ces dernières permettent la réalisation d'un premier niveau d'inférence sur applications et les documents du web.

- **La couche Ontologie** : Permet de définir une sémantique formelle et un vocabulaire commun des données utilisées dans le web, ce qui évite tous conflits de partage et d'intégration de ces données. Les ontologies permettent la réalisation de raisonnements plus élaborés sur les ressources.

- **La couche logique** : cette couche offre un ensemble de langages qui permettent l'expression des règles au niveau des ontologies, ces derniers favorisent la déduction de nouveaux faits à partir des faits existants.

- **La couche preuve** : elle fournit des moyens pour démontrer la validité des inférences données par les agents.

- **La couche confiance** : d'assurer une crédibilité aux résultats délivrés par les agents, à travers des techniques de sécurité telles que la cryptographie des messages, et l'ajout des signatures électroniques.

VI. Le web service

- Ils sont accessibles via le web par des protocoles bien connus
- Ils sont décrits à partir de XML, ils interagissent via XML
- Ils sont localisables à partir de registres
- Ils sont entièrement transversaux aux plates-formes et faiblement couplés
- Ils introduisent un nouveau modèle de développement basé sur ce que l'on appelle les architectures orientées services
- Une architecture orientée se focalise sur une décomposition plus abstraite dans la résolution dirigée par les services
- Un service résout un problème donné

- Les services peuvent être combinés pour des problèmes de plus en plus complexes

Nous notons aussi que les services web (traditionnels) possèdent des extensions sémantiques SWS) qui utilisent un ensemble de standard (OWLS, WSMO, SAWSL)[5]

VII. Web 4.0

« Any sufficiently advanced technology is indistinguishable from magic » Arthur C. Clarke

VII.1 Introduction

Le web 4.0 (Web of Things) [Guinard, 2011]], est aussi nommé le « Web Symbiotique » ou le « Web Intelligent » ou encore le « Web pervasif » ou le « Web Operating System » (ou « Web OS »). Le web 4.0 laisse la place à l'imagination, certains conceptualisent aujourd'hui le Web 4.0 comme une dimension supplémentaire de la Toile. Le Web 4.0 prolonge naturellement le concept de fédération des sources de données du Web 3.0 en étendant le type de ressource à des objets ambiants qui seront connectés (par exemple des éléments d'un bâtiment ou d'une voiture…).

Avec ce nouveau concept, nous passons de tous ce qui est réel vers le virtuel par le partage d'informations, la communication et la participation massive dans les réseaux en ligne. Cette nouvelle technologie a ouvert des nouvelles perspectives concrétisées par l'ajout et le développement des nouvelles fonctionnalités.

Le web 4.0 comme il est présenté aujourd'hui pourrait restreindre notre liberté et nos chances d'évolution et d'innovation (puisqu'il ne nous présenterait que ce qui est censé nous intéresser).

VII.2 Définition

Ces deux définitions sont issues du wikipedia

- Pour Nova Spivack, patron de Radar Networks. Il définit le Web 4.0[6] comme étant « la possibilité de travailler avec des outils uniquement en ligne ».

[5] http://www.w3.org/Submission/WSMO-related/
[6] http://www.agence-csv.com/blog/web-40/

- Une autre définition vient de **Joël de Rosnay**, conseiller du président de la Cité des Sciences et de l'Industrie de la Villette, qui le qualifie comme synonyme du « cloud computing » ou informatique en nuages.

VII.3 L'architecture du web 4.0

L'architecture du web 4.0 est base sur quatre couches

- La couche d'accessibilité de l'appareil,
- La couche Findability,
- La couche partage.
- La couche composition

Comme elle est illustrée dans la figure 1 [Guinard, 2011]

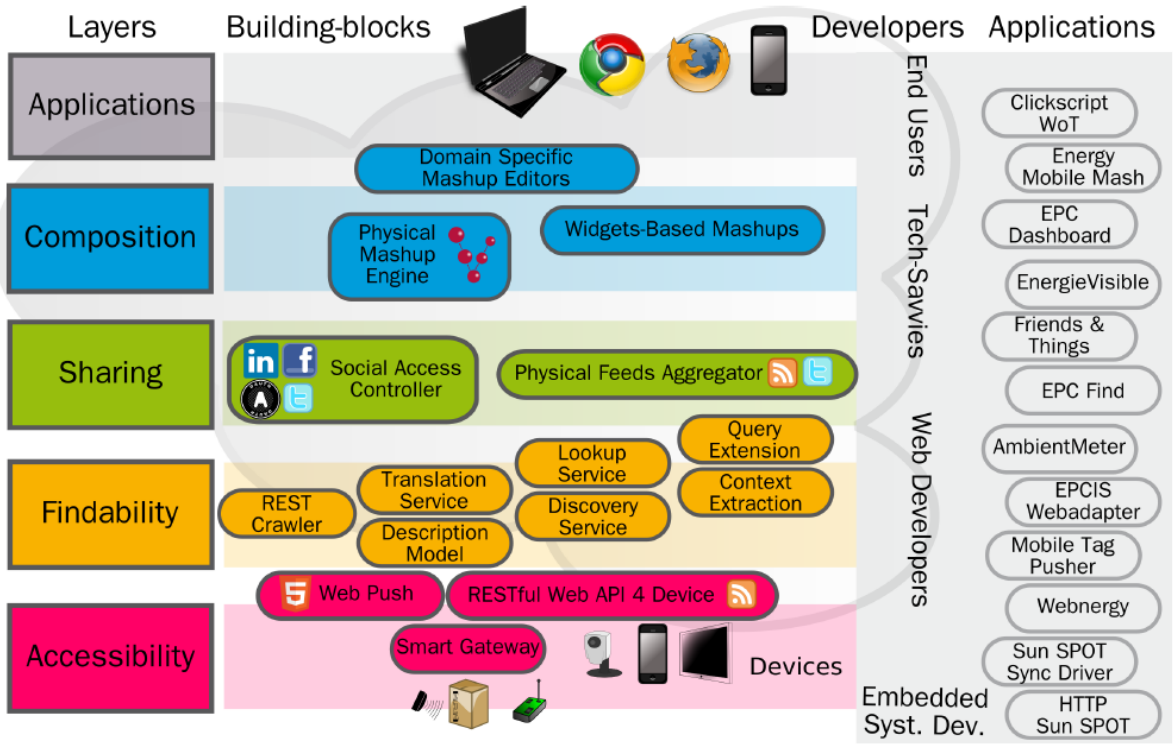

Figure II. 2: L'architecture du Web 4.0

L'objectif principal de cette architecture est de faciliter l'intégration des choses intelligentes avec les services existants sur le Web et à faciliter la création d'applications Web en utilisant des objets intelligentes (smart things).

VII.4 Les dangers du Web 4.0

Le web 4.0 peut présenter certains dangers [Guinard, 2011] dont le plus reconnu est :

- Le piratage des informations personnelles et professionnelles exposées sur la toile,
- Sécurité des sources et accès aux données,
- Capacité à analyser les comportements et à les traduire en données utiles,
- Danger d'un contrôle sur la vie privée et perte de liberté

VIII. Conclusion

Ce chapitre donne une vision globale de l'évolution de la bande, Web 1.0, web 2.0, web 3.0 et Web 4.0 (les quatre générations du web). Nous avons présenté les différentes phases de cette évolution, et chaque version possède certaines caractéristiques distinguées et fonctionnalités. Le Web 1.0 n'est que des pages statiques reliées entre eux par des liens hypertexte, l'internaute ne peut que faire la lecture de l'information, Avec l'apparition du Web 2.0 où tout internaute est contributeur du web, il est nécessaire de résoudre un nouveau type d'hétérogénéité, l'hétérogénéité sémantique. Le web sémantique(Web3.0) tente de répondre à cette hétérogénéité. Le web sémantique vise à favoriser l'appropriation du web par les utilisateurs et à réduire les incompréhensions des machines sur la signification de la connaissance. Pour cela, de nombreux standards ont été développés, tels que XML, RDF, OWL et OWRL qui sont de plus en plus actif, consolider et intégrer des contenus. L'objectif du web 4.0 est de l'innover grâces aux connections intelligentes (Operating System +cloud). Web 4.0 sera le web de lecture-écriture-exécution-concurrence.

Chapitre III : Conception et implémentation du prototype

I. Introduction

Dans les chapitres précédents, nous avons présenté les concepts de base de RI et les différents générations du web et la relations entre ces deux derniers.

De nombreux domaines de recherche sont sollicités par le développement d'Internet, notamment l'évaluation des similarités : similarité entre une requête et des documents (moteur de recherche ou système de recommandation), entre des utilisateurs ou produits (site marchand), entre des images…. Tout l'enjeu de ce plan de recherche est donc de départager de manière pertinente le nombre toujours croissant de données qui circulent sur Internet et d'en extraire des informations utiles pour répondre à divers problèmes. L'objectif fondamental de l'accès contextuel à l'information est de répondre au mieux aux besoins en information de l'utilisateur, c'est pour cette raison qu'on a conçu un système de recherche d'information basé sur les réseaux sociaux que nous avons proposé. Dans ce qui suit, nous décrivons notre approche de RI avec une proposition virtuelle pour la représentation des documents, concepts, instances et utilisateurs. Cette représentation s'inspire des réseaux sociaux, Les expérimentations que nous décrivons dans ce chapitre ont été effectuées sur des matrices crées aléatoirement. L'objectif de ces tests est de mesurer les performances et la viabilité de notre approche.

II. Présentation de la base de test

II.1 Quelques définitions

Annotation : dans le contexte des interfaces Hommes Machine [Baldonado el al. 2000] Les auteurs définissent une annotation comme un commentaire sur un objet tel que le commentateur veut qu'il soit perceptiblement distinguable de l'objet lui-même et lecteur l'interprète.

Ontologie « une ontologie formelle est spécifiée par un ensemble de noms correspondant à des concepts, et un ensemble de types de relations ordonnés selon les relations types – sous type. Les ontologies formelles sont ensuite distinguées par la façon dont les sous-types sont différentiés de leurs super-types : une ontologie axiomatisée les distingue par des axiomes et des définitions en langage formel comme certaines logiques ou certains langages informatiques traduisibles en logique ; une ontologie basée sur les prototypes les différencie par comparaison avec un membre typique, un prototype, pour chaque sous-type. Les grandes ontologies mélangent

souvent les deux approches : les axiomes et définitions sont utilisés en mathématique, physique et en science en général, les prototypes sont plus couramment utilisés pour les plantes, les animaux, et les sujets de la vie courante. » [Uschold, 1996].

- Nous avons utilisé dans nos expérimentations une ontologie issues du benchmark[7] cette dernière est codée [Raynaud et al, 2001] avec des vecteurs binaires [], afin de minimiser le temps d'exécution de la subsomption. L'ontologie se compose de 1539 concepts où chaque concept à son propre nom, numéro et code, ou le code est une représentation binaire de 28bit qui définit l'emplacement de chaque concept dans l'ontologie Exemple : 10000001010100000000000000001 est le père de tous les nœuds qui ont des bits 1 (dans la même position du père) et avec des 1 en plus par exemple ces trois concepts sont des descendants du concept précédent :
  - 10010001010100000000000001.
  - 10000001010100000001100000001.
  - 10000001010110000010000001001.

- À travers ces codes binaires on peut extraire les descendants de chaque concept en utilisant un ou logique entre le code d'un concept C_i et un code de concept candidat C_j si le résultat est le code de C_j alors C_j est descendant de C_i.
  Le père (racine) de tous les concepts a le code binaire suivant 10000000000000000000000000001.
  La taille du code =28 bits
- En ce qui concerne l'ensemble des documents que nous avons utilisé (Une collection de 100 documents) sont créés à partir d'une matrice binaire Doc-Concept générée d'une manière aléatoire, où un document peut contenir jusqu'à 1000 concepts au maximum.

  Exemple : 1 0 1 0 1 1
  1 1 1 0 0 0
  1 0 0 0 1 0

  - Doc_1 : [C_0, C_2, C_4, C_5]

[7] http://www.ws-challenge.org/.

- Doc_2 : [C_0, C_1, C_2]
- Doc_3 : [C_0, C_4]

Chaque ligne dans la matrice représente un document.

Les documents sont annotés par des tags ou encore appelés des instances. Ses derniers que nous utilisons (environ 2000 instances) ils sont en origine juste des concepts, où chaque instance appartient un concept quelconque, c'est-à-dire ils ont généré aléatoirement à travers une matrice binaire **Inst-Concept**, chaque ligne représente une instance donc chaque ligne contient un seul 1 et le reste est 0.

C'est possible que 02 l'instance I1 I2 appartiennent au même concept Ci , c'est pour ça on a utilisé des instances au lieu d'utilise des concepts directement.

Exemple : 0 0 0 1 0 0 0 0

0 1 0 0 0 0 0 0

0 0 0 0 0 0 1 0

0 1 0 0 0 0 0 0

- $Inst_0$: [C_3]
- $Inst_1$: [C_1]
- $Inst_2$: [C_6]
- $Inst_3$: [C_1]
-

- Les utilisateurs (100 utilisateurs) annotent chacun de leurs documents à travers les instances, tout ça par la création de deux matrices binaires, la première matrice [document, instance] ou **Inst-Doc**. elle détermine les instances utilisées pour annoter chacun des documents.
- et la deuxième c'est la matrice **Inst-User** pour déterminer les tags utilisés par chaque utilisateur, Chaque ligne représente l'utilisateur qui a utilisé cette instance, un utilisateur peut utiliser plusieurs instances mais une instance est utilisée par un seul utilisateur

- Exemple : matrice **Inst-doc :**

1 0 0 0 0 0 1

0 1 1 0 0 1 0

1 1 0 0 0 0 0

0 0 1 1 1 0 1

Dans la première ligne on annote les documents N°0 et N°6 avec l'instance N°0, et ainsi de suite :

- $Inst_0$ →A (Doc_0, Doc_6)
- $Inst_1$ →A (Doc_1, Doc_2, Doc_5)
- $Inst_2$ →A (Doc_0, Doc_1)
- $Inst_3$ →A (Doc_2, Doc_3, Doc_4, Doc_6)

Dans l'exemple précèdent, nous remarquons Inst1 et Inst3 ils ont le même concept, donc les documents 1, 2, 3, 4, 5, 6 ils sont annoté par le même concept C1 mais avec des utilisateurs diffèrent.

La deuxième matrice c'est la matrice **Inst-User.** Elle détermine les tags utilise par chaque utilisateur (les mêmes instances de la matrice une avec le même classement des instances), avec que chaque ligne représente l'utilisateur qui a utilisé cette instance donc dans chaque ligne il y a un seul 1 et les autres ce sont des 0. (C'est-à-dire un utilisateur peut utiliser plusieurs instance mais une instance est utiliser par un seul utilisateur

Exemple :

0 0 0 0 1 0 0
0 0 1 0 0 0 0
0 0 0 0 0 0 1
1 0 0 0 0 0 0

Pour la première ligne l'instance N°0 est utilisée par l'utilisateur N°4, Il peut se formuler comme suit :

- $Inst_0$ → [$User_4$]
- $Inst_1$ → [$User_2$]
- $Inst_2$ → [$User_6$]
- $Inst_3$ → [$User_0$]

On remarque que l'$User_2$ utilise $Inst_1$ pour annoter les documents Doc_1, Doc_2 et Doc_5, et l'$User_0$ utilise l'$Inst_3$ pour annoter les documents Doc_2, Doc_3, Doc_4 et Doc_6 qui sont à l'origine le même concept.

Et à la fin pour la recherche des documents on a besoin d'une requête, cette dernière est formulée avec une suite des concepts, Comme le montre l'exemple : Ex : C_{48} C_5 C_{125} C_{1502}.

III. Conception de l'application

Notre approche consiste à créer un système de recherche en prenant compte les tags de chaque documents dans le calcul de la mesure de similarité entre les documents et la requête.

III.1 Première partie

L'organigramme (Figure III.1) ci-dessous est composé de plusieurs étapes : après l'initialisation de la requête une étape d'extraction des descendant de chaque concept de la requête afin de les regrouper dans un ensemble en éliminant les redondances puis la même opérations pour les concepts de chaque documents, ensuite en appliquant la mesure de similarité [Skoutas & al, 2008] entre l'ensemble de descendants de la requête et l'ensemble de descendants de chaque document de la collection. Nous détaillons chacune de ces étapes dans la deuxième partie.

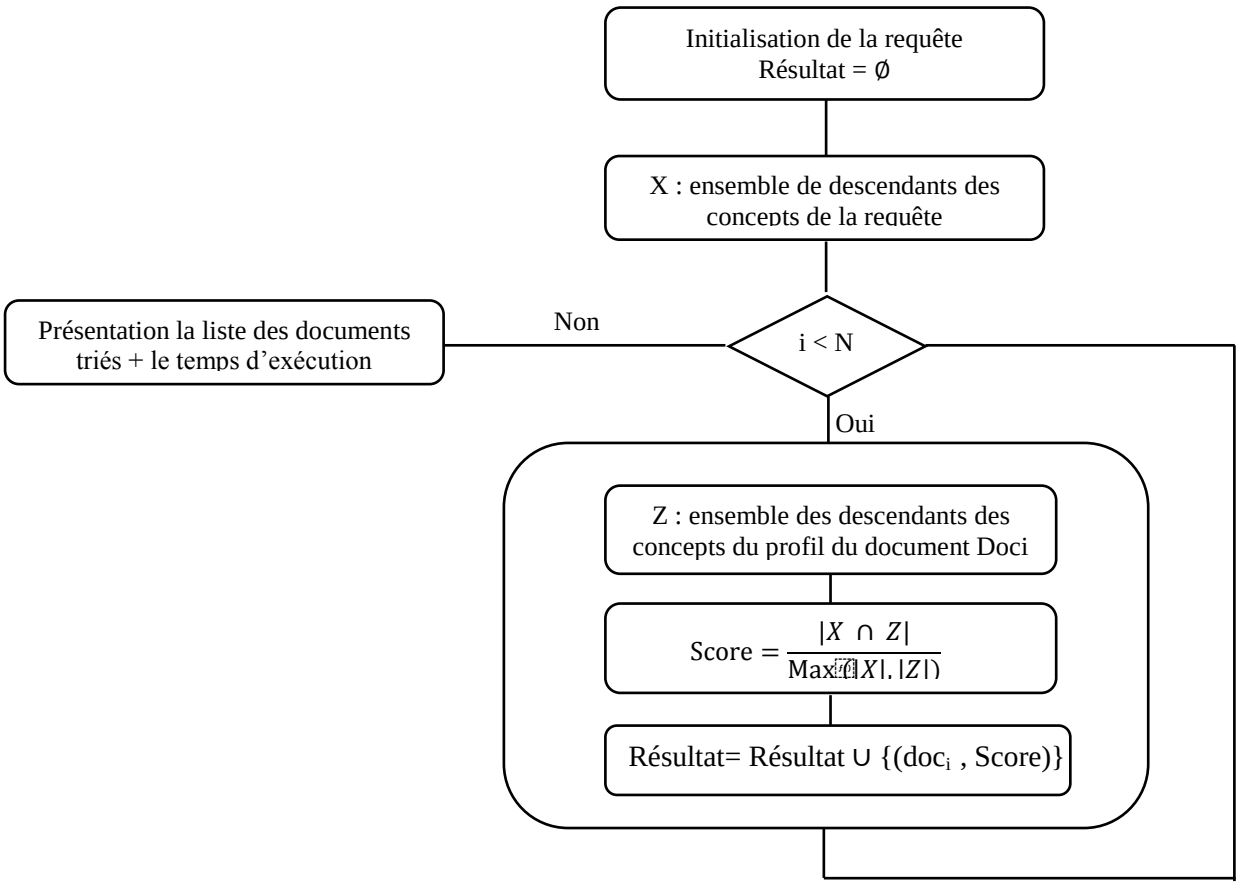

Figure III. 1 l'organigramme du système sans annotation

III.2 Deuxième partie

L'approche illustrée dans le schéma (Figure III.2) ci-dessous est composée de plusieurs étapes, que nous y allons expliquer par la suite :

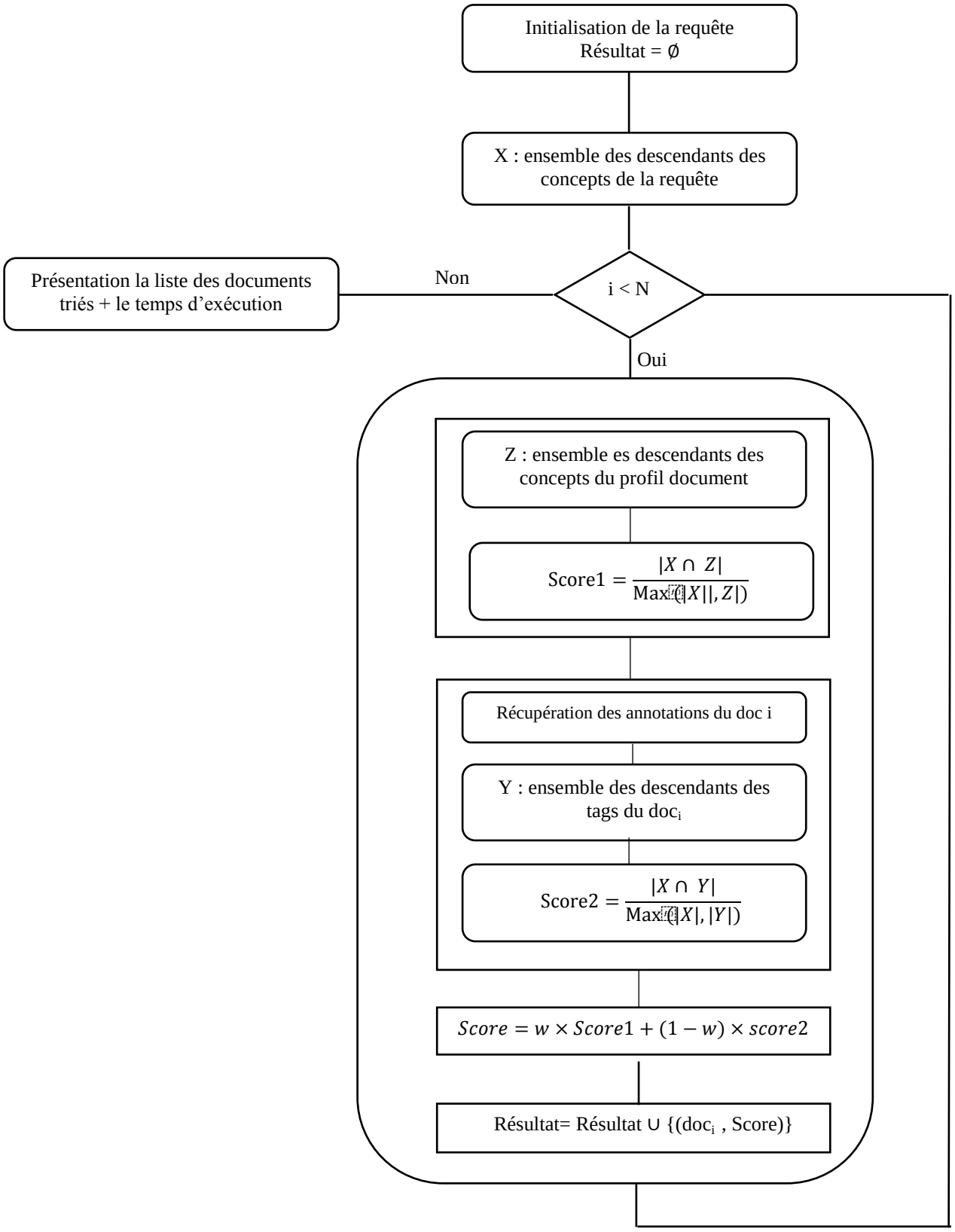

Figure III. 2 l'organigramme du système avec annotation

Etant donné que les requêtes et les documents de la base sont exprimés à l'aide des concepts de l'ontologie, Une fois la requête est initialisé une étape d'extraction des descendant de chaque concept de la requête. Puis on rentre dans la boucle pour parcourir tous les documents de la collection, tant que i (compteur) < n (nombre totale de document) si la condition est vérifié alors on fait les étapes suivant (entrer dans la boucle) sinon afficher le résultat (doc + score + temps d'exécution) une fois nous entrons dans la boucle, quatre phases nécessaires pour le résultat de la comparaison doc/requête :

- ✓ La première phase contient deux étapes :
  - la première étape est Semblable à celles de la requête (calcul des descendants de chaque élément de l'ensemble des concepts du document).
  - la deuxième étape est l'opération de calculer le score 1 : on calcule le nombre de des concepts produits à partir de l'intersection des descendant de la requête avec les descendants de document sur le maximum entre les deux.
- ✓ La deuxième phase contient trois étapes :
  - la première étape c'est la récupération des annotations ou les tags du document courant i, ces tags sont issus de tous les utilisateurs.
  - une deuxième étape comme la première étape de la phase une
  - la troisième étape c'est le calcul du score 2 qui est similaires au premier score, mais on remplace les documents par les annotations (les descendants des annotations)
- ✓ la troisième phase c'est le calcul du score final représenter par la formule suivante : Score final = w (score1) + (1-w) score2. Avec w une valeur entre 0 et 1
- ✓ et à la fin et après chaque itération on ajoute le document avec leur score dans le résultat.

Une fois la boucle est terminée il nous reste que afficher le résultat de la requête c'est-à-dire les documents triée avec le score de chaque uns plus le temps d'exécution.

Au final, basé sur le principe de folksonomie le schéma suivant représente les relations entre les différents modules du SRI :

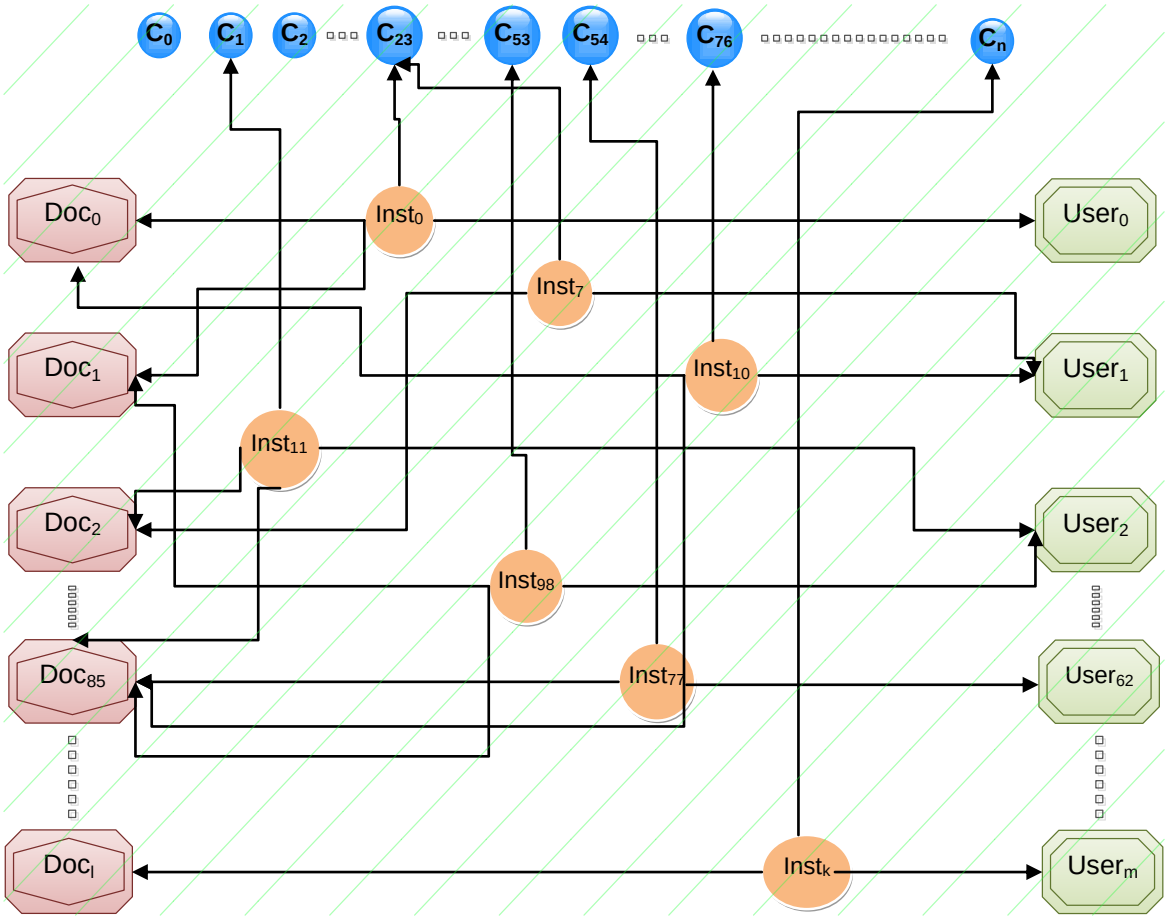

Figure III. 3 Graphe de folksonomie du système

IV. Présentation du prototype

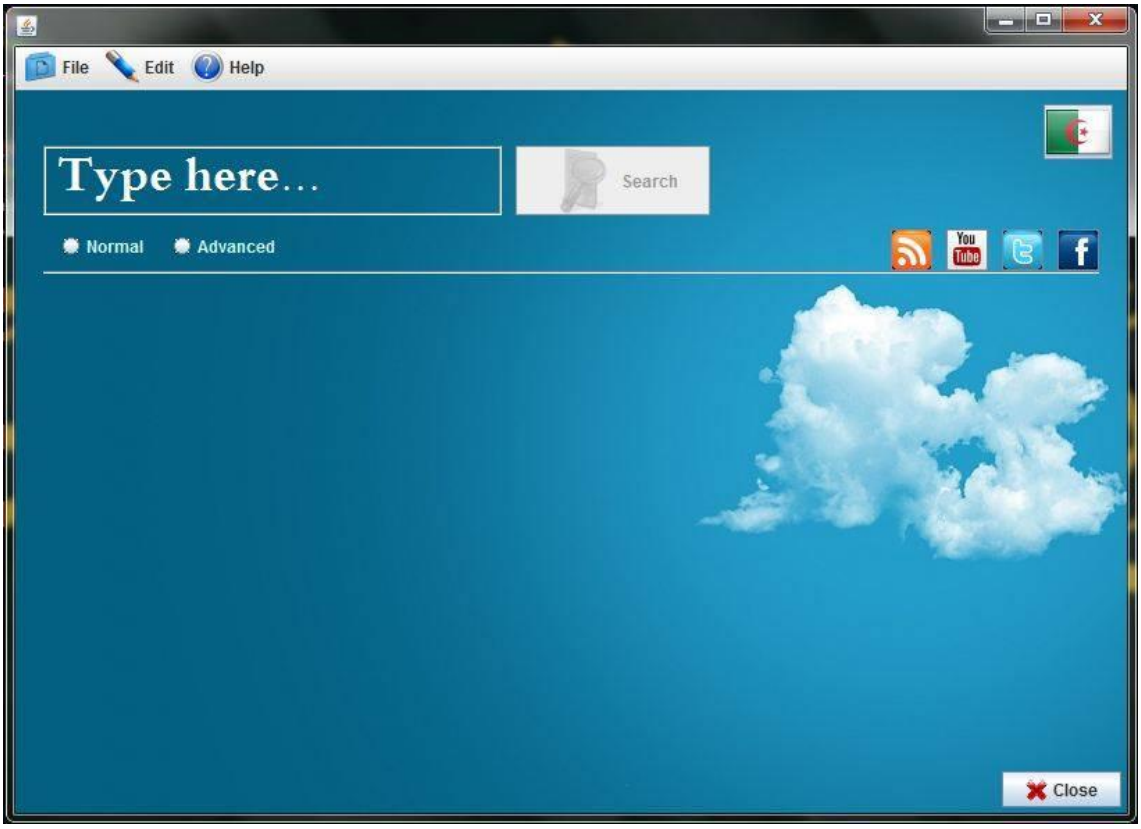

Figure III. 5 Fenêtre principale du prototype

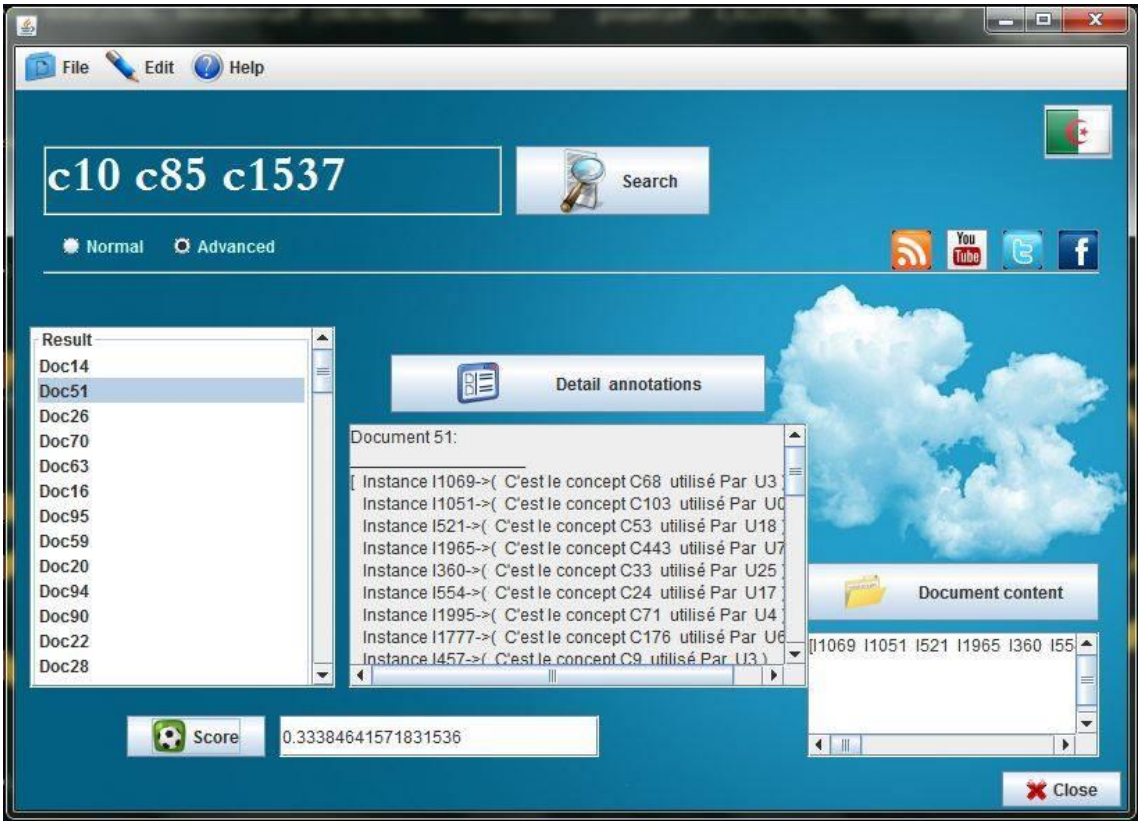

Figure III. 4 : Résultat d'une requête

Les 2 figures ci-dessus représentent l'interface du système. La première représente la fenêtre principale du système, la deuxième l'affichage du résultat retourné par le système. Pour avoir une idée comment le système fonctionne, une fois on tape la requête nous avons deux modes de recherche : normal ou avancé. Le mode normal c'est la recherche sans annotation, le deuxième mode c'est la recherche avec annotation, une fois on sélectionne un choix parmi ces deux derniers, on peut lancer l'opération de la recherche en cliquant sur le bouton « search », ensuite le système va afficher le résultat en bas à gauche des documents triés, pour voir les informations concernant un tel document, il suffit de le choisir parmi la liste et de cliquer sur le bouton « Document content » pour voir le contenu du document, le bouton « Score » pour afficher le score, le bouton « Detail annotations » pour voir tous les détails des annotations (les utilisateurs qui ont annoté ,les instances utilisé pour annoter le document…).

V. Expérimentation et résultat

V.1 Outil de travail

V.1.1 Netbeans

NetBeans[8] est un environnement de développement intégré (EDI), placé en open source par Sun en juin 2000 sous licence CDDL et GPLv2 (Common Development and Distribution License). En plus de Java, NetBeans permet également de supporter différents autres langages, comme Python, C, C++, JavaScript, XML, Ruby, PHP et HTML. Il comprend toutes les caractéristiques d'un IDE moderne (éditeur en couleur, projets multi-langage, refactoring, éditeur graphique d'interfaces et de pages Web).Conçu en Java, NetBeans est disponible sous Windows, Linux, Solaris (sur x86 et SPARC), Mac OS X ou sous une version indépendante des systèmes d'exploitation (requérant une machine virtuelle Java).

Un environnement Java Development Kit JDK est requis pour les développements en Java. NetBeans constitue par ailleurs une plate-forme qui permet le développement d'applications spécifiques (bibliothèque Swing (Java)). L'IDE NetBeans s'appuie sur cette plate-forme, il s'enrichit à l'aide de plugins.

[8] https://netbeans.org/

V.1.2 Edraw Max

Edraw Max[9] est un logiciel polyvalent de conception de diagrammes, avec des caractéristiques qui le rendent parfait non seulement pour éditer des diagrammes de flux dans un style très professionnel, des organigrammes, des diagrammes et graphiques des ventes, mais aussi pour réaliser des diagrammes réseaux, des plans de construction, des cartes heuristiques, des flux de données, des diagrammes de conceptions, des diagrammes UML, des diagrammes d'ingénierie en électricité, des illustrations scientifiques etc...

V.2 Résultats

Les tables ci-dessous représentent un échantillon des résultats retournés par le système.
La première colonne représente les documents triés (top10), la deuxième colonne nommée score c'est les valeurs calculées à partir de la mesure utilisée. On a lancé trois requêtes, pour chaque requête six tables, la première table pour la recherche sans annotation, le reste des tables la recherche est effectuée avec annotations, cinq valeurs du poids w (0.1, 0.3, 0.5, 0.7, 0.9) ont été testées dans cette dernière.

[9] http://www.edrawsoft.com/

La Requête est : c1 c10 c156 c1302	
Top 10	**Score**
Doc 73	0.03805175
Doc 79	0.03790613
Doc96	0.0309789
Doc2	0.03097345
Doc 12	0.03066037
Doc 34	0.03053435
Doc 63	0.02918069
Doc 26	0.02895322
Doc 45	0.02876712
Doc 95	0.02850877

Table III.1 : Résultat 1 de la première requête

La Requête est : c1 c10 c156 c1302, w=0.1	
Top 10	**Score**
Doc 73	0.00380517
Doc 79	0.00379061
Doc96	0.00309789
Doc2	0.00309734
Doc 12	0.00306603
Doc 34	0.00305343
Doc 63	0.00291806
Doc 26	0.00289532
Doc 45	0.00287671
Doc 95	0.00285087

Table III.2 : Résultat 2 de la première requête

La Requête est : c1 c10 c156 c1302, w=0.3	
Top 10	**Score**
Doc 73	0.01141552
Doc 79	0.01137184
Doc96	0.00929368
Doc2	0.00929203
Doc 12	0.00919811
Doc 34	0.00916030
Doc 63	0.00875420
Doc 26	0.00868596
Doc 45	0.00863013
Doc 95	0.00855263

Table III.3 : Résultat 3 de la première requête

La Requête est : c1 c10 c156 c1302, w=0.5	
Top 10	**Score**
Doc 73	0.01902587
Doc 79	0.01895106
Doc96	0.01548946
Doc2	0.015486725
Doc 12	0.01533018
Doc 34	0.01526717
Doc 63	0.01459034
Doc 26	0.01447661
Doc 45	0.01438356
Doc 95	0.01425438

Table III.4 : Résultat 4 de la première requête

La Requête est : c1 c10 c156 c1302, w=0.7	
Top 10	**Score**
Doc 73	0.03817468
Doc96	0.03322371
Doc2	0.03321987
Doc 12	0.03300072
Doc 34	0.03291250
Doc 26	0.03180572
Doc 45	0.03167544
Doc 84	0.03117168
Doc 31	0.03098290
Doc 29	0.03087102

Table III. 6 : Résultat 5 de la première requête

La Requête est : c1 c10 c156 c1302, w=0.9	
Top 10	**Score**
Doc 73	0.03424657
Doc 79	0.03411552
Doc96	0.02788104
Doc2	0.02787610
Doc 12	0.02759433
Doc 34	0.02748091
Doc 63	0.02626262
Doc 26	0.02605790
Doc 45	0.02589041
Doc 95	0.02565789

Table III. 5 : Résultat 6 de la première requête

La Requête est : c362 c413 c921 c1524	
Top 10	**Score**
Doc 45	0.10821917
Doc 39	0.09696969
Doc 12	0.09433962
Doc 67	0.09175377
Doc 74	0.08916478
Doc 14	0.08893956
Doc 63	0.08866442
Doc 51	0.08836689
Doc 70	0.08773903
Doc 95	0.08771929

Table III. 9 : Résultat 1 de la deuxième requête

La Requête est : c362 c413 c921 c1524, w= 0.1	
Top 10	**Score**
Doc 45	0.01082191
Doc 39	0.00969696
Doc 12	0.00943396
Doc 67	0.00917537
Doc 74	0.00891647
Doc 14	0.00889395
Doc 63	0.00886644
Doc 51	0.00883668
Doc 70	0.00877390
Doc 95	0.00877192

Table III. 8 : Résultat 2 de la deuxième requête

La Requête est : c362 c413 c921 c1524, w=0.3	
Top 10	**Score**
Doc 45	0.03246575
Doc 39	0.02909090
Doc 12	0.02830188
Doc 67	0.02752613
Doc 74	0.02674943
Doc 14	0.02668187
Doc 63	0.02659932
Doc 51	0.02651006
Doc 70	0.02632170
Doc 95	0.02631578

Table III. 7 : Résultat 3 de la deuxième requête

La Requête est : c362 c413 c921 c1524, w=0.5	
Top 10	**Score**
Doc 45	0.05410958
Doc 39	0.04848484
Doc 12	0.04716981
Doc 67	0.04587688
Doc 74	0.04458239
Doc 14	0.04446978
Doc 63	0.04433221
Doc 51	0.04418344
Doc 70	0.04386951
Doc 95	0.04385964

Table III. 11 : Résultat 4 de la deuxième requête

La Requête est : c362 c413 c921 c1524, w=0.7	
Top 10	**Score**
Doc 45	0.07575342
Doc 39	0.06787878
Doc 12	0.06603773
Doc 67	0.06422764
Doc 74	0.06241534
Doc 14	0.06225769
Doc 63	0.06206509
Doc 51	0.06185682
Doc 70	0.06141732
Doc 95	0.06140350

Table III. 12 : Résultat 5 de la deuxième requête

La Requête est : c362 c413 c921 c1524w=0.9	
Top 10	**Score**
Doc 45	0.09739726
Doc 39	0.08727272
Doc 12	0.08490566
Doc 67	0.08257839
Doc 74	0.08024830
Doc 14	0.08004561
Doc 63	0.07979797
Doc 51	0.07953020
Doc 70	0.07896512
Doc 95	0.07894736

Table III. 10 : Résultat 6 de la deuxième requête

La Requête est : c85 c185 c1196	
Top 10	**Score**
Doc 7	0.10469314
Doc 49	0.09305993
Doc 92	0.08936825
Doc 72	0.08862629
Doc 2	0.08554572
Doc 81	0.07793923
Doc 20	0.07464607
Doc 60	0.07301980
Doc 99	0.07067137
Doc 12	0.06957547

Table III. 15 : Résultat 1 de la troisième requête

La Requête est : c85 c185 c1196, w=0.1	
Top 10	**Score**
Doc 7	0.01046931
Doc 49	0.00930599
Doc 92	0.00893682
Doc 72	0.00886262
Doc 2	0.00855457
Doc 81	0.00779392
Doc 20	0.00746460
Doc 60	0.00730198
Doc 99	0.00706713
Doc 12	0.00695754

Table III. 14 ; Résultat 2 de la troisième requête

La Requête est : c85 c185 c1196, w=0.3	
Top 10	**Score**
Doc 7	0.03140794
Doc 49	0.02791798
Doc 92	0.02681047
Doc 72	0.02658788
Doc 2	0.02566371
Doc 81	0.02338177
Doc 20	0.02239382
Doc 60	0.02190594
Doc 99	0.02120141
Doc 12	0.02087264

Table III. 13 : Résultat 3 de la troisième requête

La Requête est : c85 c185 c1196, w=0.5	
Top 10	**Score**
Doc 7	0.05234657
Doc 49	0.04652996
Doc 92	0.04468412
Doc 72	0.04431314
Doc 2	0.04277286
Doc 81	0.03896961
Doc 20	0.03732303
Doc 60	0.03650990
Doc 99	0.03533568
Doc 12	0.03478773

Table III. 16 : Résultat 4 de la troisième requête

La Requête est : c85 c185 c1196, w=0.7	
Top 10	**Score**
Doc 7	0.07328519
Doc 49	0.06514195
Doc 92	0.06255778
Doc 72	0.06203840
Doc 2	0.05988200
Doc 81	0.05455746
Doc 20	0.05225225
Doc 60	0.05111386
Doc 99	0.04946996
Doc 12	0.04870283

Table III. 17 : Résultat 5 de la troisième requête

La Requête est : c85 c185 c1196, w=0.9	
Top 10	**Score**
Doc 7	0.09589049
Doc 49	0.08542060
Doc 92	0.08043143
Doc 2	0.07699115
Doc 81	0.07181197
Doc 20	0.06884813
Doc 60	0.06571782
Doc 14	0.06324021
Doc 12	0.06261792
Doc 70	0.06240907

Table III. 18 : Résultat 6 de la troisième requête

V.3 Discussion

Nous remarquons dans la majorité des résultats que les documents retournés (soit dans la recherche sans ou avec annotation) sont les même, en plus de ça le score de la première est supérieur à celui de la recherche avec annotation

Ceci est justifié comme suit : puisque nous utilisons un benchmark synthétique (la base est généré aléatoirement) alors les profils des documents, les annotations sociales et même les requêtes tendent à être des matrices creuses, et bien sûr lorsqu'on applique des mesures de similarités sur des matrices creuses, les scores seront très petits (proche de 0), ceci est très clair pour la recherche avec annotation.

VI. Conclusion

Dans ce chapitre nous avons présenté l'approche qu'on a proposée, ainsi que les différentes étapes (Figure III.2) de cette dernière qui ont étés suivies au cours de la réalisation du système de recherche.

Les tests effectués ont l'objectif est d'évaluer la validité de notre solution. Nous notons que l'utilisation d'un corpus réel permet de mettre en évidence l'apport du contexte social dans les performances du système de recherche d'information.

Conclusion générale

I. Conclusion

Le travail présenté dans ce mémoire s'inscrit dans le contexte général de la recherche d'information. La notion de similarité est centrale en RI. En effet, le but d'un SRI est de trouver les documents similaires à une requête formulée par un utilisateur. Le SRI doit être en mesure de comparer les documents disponibles et la requête. Cette comparaison se fait le plus souvent sur les concepts communs aux documents et à la requête. Dans l'approche proposée l'utilisation seule des concepts communs pour la mise en correspondance des documents et de la requête est insuffisante, c'est-à-dire qu'elle ne permet pas forcément de trouver tous les documents pertinents par rapport à une requête. Notre point de vue est qu'il est possible qu'un document soit pertinent pour une requête donnée sans pour autant contenir les concepts de la requête. Ce dernier peut être retourné par le système qu'on a réalisé grâce à l'intégration des tags(les annotations des utilisateurs) de chaque document dans la similarité entre le document et la requête. Ce système a la spécificité d'intégrer la pertinence thématique et la pertinence social (tags à partir des réseaux sociaux).ceci est réalisé grâce à la combinaison de deux scores, le premier score mesure la proximité entre les concepts des documents et la requête, et l'autre entre les tags du document et la requête.

II. Perspectives

Ce travail, comme tout travail de mémoire, pose de nombreuses questions et donne lieu à de nombreuses perspectives qui sont découlent de nos travaux, parmi lesquelles :

A court terme nous prévoyons :

- Utiliser un corpus réel et évaluer la performance du SRI proposé
- Améliorer le temps d'exécution d'une requête c'est-à-dire rendre le système plus rapide pendant l'opération de la recherche
- enrichir le contexte informationnel social de l'utilisateur avec les annotations de son voisinage par l'exploitation de la relation sociale.
- Nous souhaitons aussi explorer les différentes méthodes de combinaison de score social pour pouvoir comparer et approfondir les résultats obtenus.
- utiliser un thesaurus linguistique, en l'occurrence WordNet

Références Bibliographiques

[Aliane et al, 2007] H. Aliane, Z. Alimazighi, R. O. Boughacha, T. Djelliout, Un Système de reformulation de requêtes pour la recherche d'information, Centre de Recherche sur l'Information Scientifique et Technique, Alger, Algérie.2007.

[Asselin, 2008] Christophe Asslin, Le Web 2.0 pour la veille et la recherche d'information, Digimind white paper, 2008.

[Audet, 2010] Lucie Audet, WIKIS, BLOGUES ET WEB 2.0 Opportunités et impacts pour la formation à distance, Document préparé pour le Réseau d'enseignement francophone à distance du Canada, 2010.

[Baldonado & al, 2000] Baldonado M, Cousins S ,Notable at the intersection of Annotation and handheld technologies, proceeding of HUC ,conference, Bristol 2000.

[Baziz, 2005] Mustapha Baziz, indexation conceptuelle guidée par ontologie pour la recherche d'information, thèse de doctorat, Université Paul Sabatier, 2005.

[Benaouicha, 2009] Mohamed Benaouicha, Une approche algébrique pour la recherche d'information structurée, thèse de doctorat, Université Paul Sabatier, 2009.

[Binh, 2008] Nguyen Thanh Binh, Mashing-up Forge Logicielle, rapport, Institut de la Francophonie pour l'Informatique**,** 2008.

[Brini, 2005] Asma Hedia Brini, Un Modèle de Recherche d'Information basé sur les Réseaux Possibilistes, thèse de doctorat, Université Paul Sabatier, 2005.

[Cai & al, 2003] Jun Cai, Vladimir Eske, Xueqiang Wang, Semantic Web & ontologies, 2003.

[Caron, 2004] Yves Caron, Contribution de la loi de Zipf à l'analyse d'images, thèse de doctorat, Université François Rabelais de Tours 2004.

[Chaimbault, 2007] Thomas Chaimbault, Web 2.0 : l'avenir du web ?, article, école nationale supérieure des sciences de l'information et des bibliothèques, 2007.

[Chebili, 2011] Hicham Chebili, Agrégation des résultats dans la recherche d'information Semi-Structurée, Mémoire de Magister, Ecole Nationale Supérieure en Informatique, Informatique, 141 p. 2011.

[Christopher & al, 2009] Christopher D. Manning, Prabhakar Raghavan, Hinrich Schütze, An Introduction to Information Retrieval, Online edition (c)2009 Cambridge UP.2009

[Crepel, 2011] Tagging et folksonomies : pragmatique de l'orientation sur le Web, thèse de doctorat, Université Rennes 2, 2011.

[Daoud, 2009] Mariam Daoud, Accès personnalisé à l'information : approche basée sur l'utilisation d'un profil utilisateur sémantique dérivé d'une ontologie de domaines à travers l'historique des sessions de recherche, thèse de doctorat, Université Paul Sabatier, 2009.

[Guinard, 2011] Dominique Guinard, A Web of Things Application Architecture - Integrating the Real-World into the Web, thèse de doctorat, 2011.

[Hammache, 2013] Arezki Hammach, Recherche d'information : un modèle de langue combinant mots simples et mots composés, thèse de doctorat, Université Tizi-Ouzou ,2013.

[Harrathi, 2010] Rami Harrathi, Recherche d'information conceptuelle dans les documents semi-structurés, thèse de doctorat, Université Lyon, 2010

[Hlaoua, 2007] Lobna Hlaoua, Reformulation de Requêtes par Réinjection de Pertinence dans les Documents Semi-Structurées, thèse de doctorat, Université Paul Sabatier, 2007.

[Ho, 2004] Ho Bao-Quac, Vers une indexation structurée basée sur des syntagmes nominaux, thèse de doctorat, Université Joseph Fourier-Grenoble I, 2004.

[Karbasi, 2007] Soheila Karbasi, Pondération des termes en Recherche d'Information :

Modèle de pondération basé sur le rang des termes dans les documents, thèse de doctorat, Université Paul Sabatier, 2007.

[Kompaoré, 2008] Nongdo Désiré Yawbsom Kompaoré, Fusion de systèmes et analyse des caractéristiques linguistiques des requêtes : vers un processus de RI adaptatif, thèse de doctorat, Université Paul Sabatier, 2008.

[Le Targat, 2005] Gaëlle Le Targat, Langages classificatoires et recherche d'information sur les portails d'entreprise : quels apports pour les utilisateurs ?, thèse de doctorat, 2005.

[loblet & al, 2002] Philippe Laublet, Chantal Reynaud, Jean Charlet, Sur quelques aspects du Web sémantique, 2002.

[Luong, 2012] The Nhan Luong, Modélisation centrée utilisateur final appliquée à la conception d'applications interactives en géographie : une démarche basée sur les contenus et les usages, thèse de doctorat, Université de Pau et des Pays de l'Adour, 2012.

[Marini, 2010] Jean-Luc Marini, Capitalisation d'expériences pour l'indexation et la recherche d'information dans le domaine de la Gestion Electronique de Documents, thèse de doctorat, 2010.

[Mathilde, 2009] Giraud Mathilde, Les réseaux sociaux peuvent-ils devenir un nouvel outil de marketing pour l'entreprise ?, mémoire fin d'étude, 2009.

[Nassr, 2002] Nawel Nassr, Croisement de langues en recherche d'information : traduction et désambiguïsation de requêtes, thèse de doctorat, Université Paul Sabatier, 2002.

[O'Reilly & al, 2009] Tim O'Reilly et John Battelle. « Web 2.0 Summit », 2009

[Pruski, 2009] Cédric Pruski, Une approche adaptative pour la recherche d'information sur le Web, thèse de doctorat, Université du Luxembourg, 2009.

[Quoniam & al, 2009] Luc Quoniam, Arnaud Lucien, Du web 2.0 à l'intelligence, compétitive 2.0, Les Cahiers du numérique (Vol. 5) Pages : 196, avril 2009.

[Raynaud et al, 2001] Olivier Raynaud, Eric Thierr ,A Quasi Optimal Bit-Vector Encoding of Tree Hierarchies. Application to Efficient Type Inclusion Tests, ECOOP '01 Proceedings of the 15th European Conference on Object-Oriented Programming,Pages 165-180 2001.

[Ressad-Bouidghaen, 2011] Ourdia Ressad-Bouidghaen, Accès contextuel à l'information dans un environnement mobile : approche basée sur l'utilisation d'un profil situationnel de l'utilisateur et d'un profil de localisation des requêtes, thèse de doctorat, Université Paul Sabatier, 2009.

[Rupert, 2009] Maya Rupert, Coévolution d'organisations sociales et spatiales dans les systèmes multi-agents : application aux systèmes de tagging collaboratifs, thèse de doctorat, Université de Lyon, 2009.

[Salton & al, 1989] Gerard Salton. Addison-Wesley, Automatic text processing: the transformation, analysis, and retrieval of information by computer. Reading, MA, 1989.

[Seilles, 2012] Antoine Seilles, Structuration de débats en ligne à l'aide d'Annotations socio-sémantiques, thèse de doctorat, Université Montpellier II, 2012.

[Skoutas & al, 2008] D Skoutas, D Sacharidis, V Kantere, T Sellis, Efficient semantic web service discovery in centralized and p2p environments The Semantic Web-ISWC 2008, 583-598.

[Trudeau, 2010] Lyn Trudeau, le Web en évolution constante, 5 décembre 2010.

[Uschold, 1996] M. Uschold. Converting an Informal Ontology into Ontolingua: Some Experiences. A slightly abridged version of this paper appears in the Proceedings of the Workshop on Ontological Engineering held in conjunction with ECAI 96, Budapest, 1996.

[W3C, 1998] http://www.w3.org/DesignIssues/Semantic.html consulté le 15/04/2014.

[Yaël, 2009] Yaël Champclaux, Un modèle de recherche d'information basé sur les graphes et les similarités structurelles pour l'amélioration du processus de recherche d'information, Université Toulouse, thèse de doctorat, 2009.

[Zammar, 2012] Nisrine Zammar, Réseaux Sociaux Numériques : Essai de catégorisation et de cartographie des controverses, thèse de doctorat, Université Rennes 2, 2012.

Listes des figures

Listes des tables

Liste des abréviations

- C : Concpet
- D_j ou Doc : document
- Inst : Instance
- P : Précision
- Q : Query (requte)
- R : Rappel
- RI : Recherche d'information
- RS : Réseaux sociaux
- RSV : Retrieval Status Values
- SRI : Système de recherche d'information