Université Saad DAHLAB - Blida 1

Faculté des sciences

Département d'Informatique

Mémoire présenté par :

MllesBachebacheMarwa et Bounekhla Meriem

Pour l'obtention du diplôme de Master

Domaine : Mathématique et Informatique

Filière : Informatique

Spécialité : Ingénierie des Logiciels

Sujet :

Proposition d'une méthode d'indexation, à la volée, des documents textuels à l'aide des Folksonomies

Soutenu le : 29/09/1019, devant le jury composé de :

Mme. Oukid Saliha	Présidente	Université de Blida 1
Mme. Ghebghoub Yasmine	Examinatrice	Université de Blida 1
Mme. Mezzi Melyara	Promotrice	Université de Blida 1

Résumé

Habituellement, les Systèmes de recherche d'Information (SRI) n'effectuent pas la recherche directement sur les collections de données (datasets)car ce traitement est contraignant en terme de temps et d'espace mémoire, mais font plutôt recours à ce que l'on appelle des indexes, du genre que nous trouvons dans la fin des livres. De ce fait, à travers les années, l'indexation est devenue sans conteste, la tâche la plus onéreuse parmi toutes les autres tâches de RI. Le but principal de l'indexation est de créer une représentation des documents présents dans le corpus de documents global de façon à automatiser les traitements de ces derniers et faciliter leur appariement avec les requêtes des utilisateurs. Cette représentation, plus allégée, est alors, enregistrée sous une structure appropriée pour faciliter l'accès à l'information et le renvoi de documents pertinents en un minimum de temps.

Dans ce projet, nous nous intéressons à l'étude de l'utilisation des Floksonomies comme moyen rapide et personnalisé de représenter et d'indexer l'information car considérées comme l'alternative moderne aux Ontologies et ressources lexicales traditionnellement utilisées dans la RI. D'autres part, le but est d'analyser si cette pratique peut améliorer l'efficacité d'un système de RI en permettant aux documents d'être indexés et donc retrouvés en temps réel (c.-à-d. : dès leur mise en ligne), et ce en se basant sur le contexte social.

Mots clés : Système de Recherche d'Information, Indexation, Sensibilité au contexte, Folksonomies, Ontologies et web sémantique, Traitement Automatique de la Langue.

Abstract

Usually, Information Search Systems (IRS) do not search directly on data collections (datasets) because this processing is constraining in terms of time and memory space, but use what are called indexes, of the kind we find at the end of books. As a result, over the years, indexation has undoubtedly become the most expensive of all other IR tasks. The main purpose of indexing is to create a representation of the documents present in the global corpus of documents in order to automate their processing and facilitate their matching with users' requests. This lighter representation is then recorded under an appropriate structure to facilitate access to information and the return of relevant documents in a minimum of time.

In this project, we are interested in studying the use of Floksonomies as a fast and personalized way to represent and index information as considered as the modern alternative to Ontologies and lexical resources traditionally used in IR. On the other hand, the aim is to analyse whether this practice can improve the effectiveness of an IR system by allowing documents to be indexed and therefore retrieved in real time (i. e. as soon as they are put online), based on the social context.

Keywords: Information Retrieval, Semantic web, Stemming, Context-modelling, Social-bookmarking, Folksonomies, Natural Language Processing.

ملخص

عادة، لا تبحث أنظمة استرجاع المعلومات (IRS) مباشرة من مجموعات البيانات لأن هذه المعالجة مقيدة من حيث الوقت ومساحة الذاكرة، ولكنها تستخدم ما نسمي الفهارس، من النوع الذي نجده في نهاية الكتب. نتيجة لذلك، أصبحت الفهرسة على مر السنين أهم مهمة بين جميع مهام IR الأخرى. الغرض الرئيسي من الفهرسة هو إنشاء تمثيل للمستندات الموجودة في مجموعة المستندات العالمية وذلك لأتمتة معالجة هذه المستندات وتسهيل مطابقتها لطلبات المستخدمين. ثم يتم تسجيل هذا التمثيل الأخف في ظل هيكل مناسب لتسهيل الوصول إلى المعلومات وإعادة المستندات ذات الصلة في فترة زمنية لا تقل عن الحد الأدنى.

في هذا المشروع، نحن مهتمون بدراسة استخدامFloksonomies كوسيلة سريعة وشخصية لتمثيلوفهرسة المعلومات باعتبارها البديل الحديث للأنطولوجيا والموارد المعجمية المستخدمة تقليديًا فيIR. من ناحية أخرى، يتمثل الهدف في تحليل ما إذا كانت هذه الممارسة يمكنها تحسين كفاءة نظام IR عن طريق السماح بفهرسة الوثائق وبالتالي استرجاعها في الوقت الفعلي (أي بمجرد طرحها عبر الإنترنت)، بناءً على السياق الاجتماعي.

كلمات البحث: استرجاعالمعلومات, ويبالدلالي, الجذعية, السياقالمنمذج, المفضلاتالاجتماعية, معالجةاللغةالتلقائي, Folksonomies

Dédicaces

C’est avec un immense plaisir que je dédie ce travail

A ma mère qui est toute ma vie et tout ce que j’ai de plus cherau monde, en témoignage de ma reconnaissance infinie pour les nombreuxsacrificesqu’elle n'a cessé de déployer pour moi et dont je serais à jamaisredevable.
Que dieu la garde et la procure la santé et le bonheur.

A mon chère papa (ربي يرحمه) qui était toujours avec moi et m’a encouragé,
Il était et restera mon pouvoir et ma force dans la vie.

Qu'ils trouvent en ce travail la preuve de mon éternel amour et ma reconnaissance envers eux.

Ainsi qu’à mon frères bien aimé Mohamed qui est ma fierté et à la place de mon père et sa femme Rima, mes 2 sœurs Chahra et Fella et ses enfants aussi mes 2 beaux-frères Ahmed et Amine, en témoignage de ma grande affection pour eux.

Puisse dieu les protéger.

Merci beaucoup

Et enfin, je dédie ce travail à mes cousines Zineb et Aida et tous mes amis surtout Amira, Fatima, Abir, fatma Zohra, khadidja, Mahdi, Réda, Hichem qui ont toujours été là à mes cotés

Sans oublier mon binôme Meriem (mon acolyte dans cette épreuve) ainsi que toute sa famille.

Toute la lutte que nous pensions avoir été vaine, toutes les erreurs, une vie peut contenir.... finissent tous par s'en aller.

Mlle Bachebache Marwa

Dédicaces

Je dédié ce travail

A mes chers parents ma mère et mon père pour leur patience, leur amour, leur soutien et leur encouragement.

A mes sœurs Amina et Khadidja ainsi que sa fille Lydia, A mon frère Mohamed et mon beau-frère Abdelhak.

A ma chère cousine Nadjet et mes copines Rim, Meriem et Nesrine.

Sans oublier mon amie Marwa (mon acolyte dans cette épreuve) ainsi que toute sa famille.

Mlle Bounekhla Meriem

Remerciements

Tout d’abord, nous tenons à rendre grâce à dieu tout puissant pour nous avoir donné le courage et la détermination nécessaires pour finaliser ce travail et le mener à terme.

En second lieu et puisque une seule main ne lie pas un fagot de bois, Nous tenons à remercier chaleureusement et affectueusement tous ceux et toutes celles qui de près ou de loin ont contribué à la réalisation de notre projet.

On ne saurait ne pas remercier encore une fois nos parents respectifs qui, par leur amour et leur affection nous ont permis d’arriver là où nous sommes aujourd’hui.

Nous tenons également à remercier Mme Melyara qui a endossé son rôle de promotrice de la meilleure façon qui soit. Nous retiendrons son aide précieuse, ces conseils avisés, ces idées riches mais aussi sa sympathie et ses encouragements. Nous prions Allah de lui rendre grâce pour avoir fait de notre travail avec elle, un réel honneur et grand plaisir.

Par ailleurs ; nous rendons un vibrant hommage à l’ensemble du corps professoral du département d’informatique de l’université Saâd DAHLAB de Blida qui ont contribué activement et vaillamment à notre formation pendant ces cinq dernières années.

Dédicaces spéciale à tous nos camarades de la promotion 2018/2019, plus particulièrement à nos amis Imad et Nesrine. Puisse dieu les aider à atteindre tous leurs buts.

A tous ceux et à toutes celles dont les noms n'apparaissent pas sur cette page, qu'ils demeurent convaincus, que nous ne les avons point oubliés et qu'ils soient assurés de notre profonde gratitude.

Merci.

Table des matières

Chapitre 2: L'indexation sémantique

Chapitre 3: Les folksonomies

Liste des figures

Liste des tableaux

Liste d’acronymes

Acronyme	Intitulé
RI:	Recherche d'Information
SRI:	Système de Recherche d'Information
memex:	Memory Extender
www:	World Wide Web
TF:	Terme Frequency
IDF:	Inverse Of Document Frequency
RSV:	Retrieval Status Value
TREC:	Text Retrieval Conference
NIST:	National Institute Of Standards And Technology
TRC2:	Text Research Collection
CLEF:	Conference And Labs Of The Evaluation Forum
MOAT :	Meaning Of A Tag
JFC :	Java Foundation Classes
URL :	Uniform Resource Locator
IBM :	International Business machines corporation
NISO :	National Information Standards Organisation
SECAS :	Semantically Enriched Context Aware Stemming
CAS :	Context Aware Stemmer
XML :	Extensible Markup Language
HTML :	Hypertext Markup Language

Introduction générale

1. Contexte global

La Recherche d'Information (RI) est un domaine qui s'intéresse à la structure, à l'analyse, à l'organisation, au stockage, à la recherche et à la découverte de l'information. Le défi est de pouvoir, parmi le volume important de documents disponibles, trouver ceux qui correspondent au mieux à l'attente de l'utilisateur. L'opérationnalisation de la RI est réalisée par des outils informatiques appelés Systèmes de Recherche d'Information (SRI). Ces systèmes ont pour but de mettre en correspondance une représentation du besoin de l'utilisateur avec une représentation du contenu des documents au moyen d'une fonction de correspondance. L'évaluation d'un SRI consiste à mesurer ses performances vis-à-vis du besoin de l'utilisateur, à cet effet les méthodes d'évaluation largement adoptées en RI sont basées sur un modèle qui fournit une base d'évaluation comparative de l'efficacité de différents systèmes moyennant des ressources communes. Ces ressources sont essentiellement des collections de tests, des requêtes préalablement construites, des jugements de pertinence et des métriques d'évaluation.

Au cours des dernières décennies, il y a eu des changements remarquables dans la région de la recherche d'information (IR) car une énorme quantité d'informations est de plus en plus accumulés sur le Web. La gigantesque explosion d'informations augmente la nécessité de découvrir de nouveaux outils permettant de récupérer des connaissances significatives à partir de diverses sources d'informations complexes.

Notre travail est basé sur la Folksonomie qui désigne un système de classification collaborative par les internautes. L'idée est de permettre à des utilisateurs de partager et de décrire des objets via des mots-clés (tags) librement choisis. Formellement, une Folksonomie est composée de trois ensembles : un ensemble d'utilisateurs, un ensemble de tags (ou étiquettes) et un ensemble de ressources (films, livres, sites web, photos, etc.). Les utilisateurs sont les acteurs principaux du système et contribuent au contenu par l'ajout de ressources et l'affectation de tags. Cependant, il s'avère que le

choix de tags et de ressources partagées par un utilisateur d'une folksonomie varie selon plusieurs critères : le genre, l'âge ou encore la profession de celui qui partage l'information. Cela a incité les chercheurs à proposer des systèmes de recommandation personnalisés permettant de suggérer, selon ces tags, les ressources les plus appropriées aux utilisateurs. De plus, la personnalisation tente d'aider les utilisateurs à aborder le problème de surcharge d'information. Pour atteindre cet objectif, nous envisageons l'utilisation des Folksonomies, comme nouveau moyen permettant l'accès rapide aux informations contenues dans le web, et nous proposons une approche d'indexation basée sur les tags des utilisateurs, ce qui pourrait permettre une description pertinente des documents se trouvant sur le web sans perdre le contenu sémantique des documents..

2. Problématique

Un Système de Recherche d'Informations (SRI) vise à sélectionner des documents pertinents qui répondent aux besoins d'informations de l'utilisateur exprimés par une requête textuelle. Au cours des années 1970-1980, différents modèles théoriques ont été proposés dans ce sens pour représenter d'une part des documents et des requêtes, d'autre part, pour répondre aux besoins d'information indépendamment de l'utilisateur.

Plus récemment, avec l'arrivée du Web 2.0, également appelé Web social, l'efficacité de ces modèles a été remise en question, car ils ignorent le contexte dans lequel se trouvent les informations. En effet, les méthodes traditionnelles d'indexation, trop lentes et trop complexes ont vite été dépassées devant la gigantesque explosion d'informations qui a accroît le besoin de découvrir de nouveaux outils permettant d'indexer et ainsi permettre la récupération des connaissances utiles à partir de diverses sources d'informations et ce, en temps réel dans le monde impatient et mobile dans lequel nous vivons.

Dans ce modeste travail, nous présentons une réflexion sur une méthode de pré-indexation rapide permettant à un document textuel de pouvoir être recherché et trouvé aussitôt qu'il est mis en ligne.

3. Objectifs de travail

Les environnements d'information modernes deviennent de plus en plus vastes ainsi que complexes et omniprésents, car la quantité de ressources disponibles et les informations hétérogènes croissent de manière exponentielle chaque année. Presque tous les aspects de nos vies sont affectés par l'information disponible sur Internet

Dans ce travail, nous nous concentrons sur la proposition d'une nouvelle technique pour indexer les documents textuels, aussitôt qu'ils sont mis en ligne. Ainsi, l'objectif principal à atteindre peut être formulé comme suit : Proposition d'un algorithme d'indexation à la volée sensible au contexte social qui permet de conserver l'effecacité des algorithmes d'indexation classique avec des résultats plus pertinents;

4. Organisation du mémoire

Afin d'atteindre l'objectif cité ci-dessus, notre mémoire s'articulera autour de deux parties :

4.1.1. Etat de l'art qui contiendra Trois chapitres :

Chapitre 1 : présente les concepts de base de la RI. Nous commençons par donner une définition de la RI et nous décrivons les différents modèles servant de cadre théorique pour la modélisation du processus de RI. Nous illustrons également le processus de RI en présentant les étapes d'indexation, d'interrogation et de mise en correspondance. Par la suite, nous présentons les outils de RI sur le web, les algorithmes des moteurs de recherche et l'architecture générale de ce type d'outil. Enfin, nous terminons par l'évaluation des modèles de RI à travers les mesures de performances.

Chapitre 2 : traite le web sémantique ainsi que ses principaux composants, nous revenons sur l'indexation sémantique en RIainsi que sur les bases de connaissance et leurs approches.

Chapitre 3 : traite notre partie principale qui est la folksonomie ainsi que ses types et caractéristiques. Nous y parlerons des tags et les applications utilisant ces approches. Nous conclurons ce chapitre par quelques travaux connexes des chercheurs.

4.1.2. Solution proposée

Chapitre 4 : Présente l'analyse des besoins et conception de notre solution. Nous parlerons de la démarche de modélisation utilisée illustré par un organigramme que nous avons proposé ainsi les mesures de similarité et les différents algorithmes qui composent la solution globale d'une technique d'indexation à la volée.

*Chapitre 5 : Présente l'*implémentation du système, dans lequel nous présenterons notre modèle de base de données, les techniques d'évaluation et l'environnement de développement (langages et outils utilisés). Nous conclurons par une partie tests et validation, où nous évaluerons les expérimentations du système et la qualité des résultats qu'il fournit comparée aux objectifs initiaux.

En fin, la conclusion de ce mémoire synthétisera nos principales contributions et donnera quelques perspectives à notre travail.

Chapitre I : Recherche d’information

1. Introduction

Les origines de la recherche d'information (RI) [**1**] peuvent revenir à l'époque de la seconde guerre mondiale ou des quantités massives de la documentation et des rapports sur les armes ont été produites. A cette époque, l'indexation des documents était déjà une tâche lourde. L'ampleur de cette tâche a été décrite dans la célèbre publication de Vannevar Bush au sujet de la *memex*(memory extender). Cette réalité n'a pas Changé depuis, mais elle est devenue une tâche encore plus complexe. La croissance d'Internet et le WWW (WorldWide Web) a généré d'énormes volumes d'informations.

Ces informations se trouvent justes à quelques clics de souris, mais l'accès à ces informations a créé un besoin crucial pour créer des outils d'aide à la recherche d'information pour satisfaire les besoins des utilisateurs. Le système qui fournit cette aide est généralement connu sous le nom *moteur de recherche*. Le terme moteur de recherche est considéré comme un synonyme de système de RI, basé sur un algorithme bien défini.

Ce chapitre est organisé en quatre parties : la première partie présente la définition de la recherche d'information. La deuxième partie, décrit les concepts de traitement de l'information dans le but de fournir une compréhension sur le processus de la recherche d'information. La troisième partie présente les principaux modèles classiques de recherche. Enfin, la dernière partie donne un aperçu sur le protocole et les métriques d'évaluation en recherche d'information.

1.1. La recherche d'information

La recherche d'information [**2**] est un domaine historiquement lié aux sciences de l'information et à la bibliothéconomie qui ont toujours eu le souci d'établir des représentations des documents dans le but d'en récupérer des informations à travers la construction d'index. L'informatique a permis

le développement d'outils pour traiter l'information et établir la représentation des documents au moment de leur indexation, ainsi que pour rechercher l'information. On peut aujourd'hui dire que la recherche d'information est un champ transdisciplinaire qui peut être étudié par plusieurs disciplines utilisant des approches qui devraient permettre de trouver des solutions pour améliorer son efficacité.

1.2. Définition

L'une des premières définitions pour les systèmes de recherche d'information est proposée par 'Salton ' et 'Ingwersen'.
En 1968, salton propose cette definition:*"Information retrieval is a field concerned with the structure, analysis, organization, storage, searching, and retrieval of information."*.

En 1980, Ingwerses donne définit la RI comme suit: "*Information retrieval systems address the representation, organization of, andaccess to large amounts of heterogeneous information encoded in digital format*."

En 1983, Salton a proposé une nouvelle définition et a défini le système de RI comme suit :"*An information retrieval system is an information system, that is, a systemused to store items of information that need to be processed, searched, retrieved, and disseminated to various user populations*."

Deux définitions plus récentes sont celles de Boubekeur: " *La recherche d'information est une branche de l'informatique qui s'intéresse à l'acquisition, l'organisation, le stockage, la recherche et la sélection d'information* " et de Daoud : " *La recherche d'information est une discipline de recherche qui intègre des modèles et des techniques dont le but est de faciliter l'accès à l'information pertinente pour un utilisateur ayant un besoin en information* "

Depuis 1968 les définitions partagent la même idée que la recherche d'information a pour but l'extraction d'information d'un document ou ensemble de document [**1**].

1.3. Concepts de traitement de la recherche d'information

Le processus de recherche d'information est composé de trois étapes : l'indexation, le requêtage et l'appariement. Ces étapes sont plus ou moins complexes en fonction de la tâche de recherche.

Nous présentons les principales étapes dans un processus de recherche d'information dans ce qui suit : [**3**]

1.4. Indexation

Consiste à identifier pour chaque document, les termes importants et exploiter ces termes comme un index pour que l'accession aux documents soit rapide alors on peut déduire que l'un des objets de l'indexation est de retrouver rapidement le document contenant les termes (mots clés) de la requête.

Il existe trois types d'indexation :

1.4.1. Indexation manuelle

Chaque document de la collection est analysé par un spécialiste du domaine ou un documentaliste. L'indexation manuelle assure une meilleure précision dans les documents restitués par le SRI en réponse aux requêtes des utilisateurs.

Néanmoins, cette indexation présente un certain nombre d'inconvénients liés notamment à l'effort et le prix qu'elle exige (en temps et en nombre de personnes).

De plus, cette indexation est subjective, liée au facteur humain. Différents spécialistes peuvent indexer un document avec des termes différents. Il se peut même arriver qu'un spécialiste indexe différemment un document, à différents moments. [**3**]

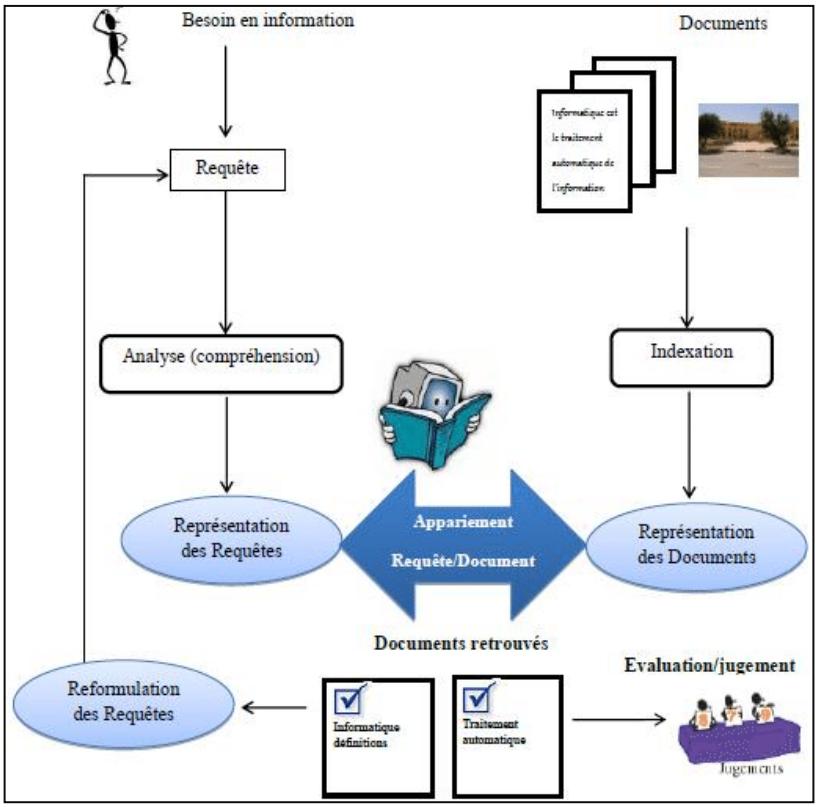

Figure 1 : Système de recherche d'information selon Baeza-Yates et Ribeiro-Neto **[4].**

1.4.2. Indexation semi-automatique

La tâche d'indexation est réalisée ici conjointement par un programme informatique et un spécialiste du domaine. Le choix final des descripteurs revient à l'indexeur humain. Dans ce type d'indexation un langage d'indexation contrôlé est généralement utilisé [**3**].

1.4.3. Indexation automatique

Dans ce cas, l'indexation est entièrement automatisée. Elle est réalisée par un programme informatique et elle passe par un ensemble d'étapes pour créer d'une façon automatique l'index. Ces étapes sont : l'analyse lexicale, l'élimination des mots vides, la normalisation (lemmatisation ou radicalisation), la sélection des descripteurs, le calcul de statistiques sur les descripteurs et les documents (fréquence d'apparition d'un descripteur dans un document et dans la collection, la taille de chaque document, etc.) et

enfin la création de l'index et éventuellement sa compression. Nous détaillons ces différentes étapes ci-dessous [**3**].

- ***Extraction des mots :*** Elle permet de convertir un texte de document en une liste de termes. Un terme est un groupe de caractères constituant un mot significatif. L'extraction des mots permet de reconnaître les espaces de séparation des mots, les chiffres, les ponctuations.
- ***Élimination des mots vides :*** Les mots vides (article, proposition, conjonction, etc.) sont des mots non significatifs dans un document, car ils ne traitent pas le sujet du document.

  On distingue deux techniques pour éliminer les mots vides :

  1. L'utilisation d'une liste préétablie de mots vides (aussi appelée anti-dictionnaire ou stopList).
  2. L'élimination des mots ayant une fréquence qui dépasse un certain seuil dans la collection.

  L'élimination des mots vides réduit la taille de l'index, ce qui améliore le temps de réponse du système. Cependant, elle peut réduire le taux de rappel en réponse à des requêtes bien spécifiques par exemple :

Anglais: the, or, a, you, I, us, ...
Français : le, la, de, des, je, tu, ...
–Des exceptions :
• US : « USA »
• A de (vitamine A)

- ***Normalisation :*** La normalisation consiste à représenter les différentes variantes d'un terme par un format unique appelé lemme ou racine. Ce qui a pour effet de réduire la taille de l'index. Plusieurs stratégies de normalisation sont utilisées : la table de correspondance, l'élimination des affixes l'algorithme de Porter, la troncature, l'utilisation des N-grammes.

  L'inconvénient majeur de cette opération est qu'elle supprime dans certains cas la sémantique des termes originaux, c'est le cas par exemple

des termes derivate/derive, activate/active, normalisés par l'algorithme de Porter.

- ***Pondération des mots :*** La pondération permet d'affecter à chaque terme d'indexation une valeur qui mesure son importance dans le document où il apparaît.

  Le pouvoir de discrimination des termes pour décrire le contenu des documents n'est pas identique pour tous les termes. Pour trouver les termes du document qui représentent le mieux son contenu sémantique, la communauté IR a défini la fonction de pondération d'un terme dans un document connue sous la forme de Tf.Idf, qui est reprise dans différentes versions par la majorité des SRI. On y distingue :

  - Tf (termfrequency) : cette mesure est proportionnelle à la fréquence du terme dans le document. L'idée sous-jacente est que plus un terme est fréquent dans un document, plus il est important dans la description de ce document. Le Tf est souvent exprimé selon l'une des déclinaisons suivantes :
    - Tf : utilisation brute
    - $0.5 + 0.5(Tf/\max(Tf))$
  - Idf (Inverse of Document Frequency) : mesure l'importance d'un terme dans toute la collection. L'idée sous-jacente est que les termes qui apparaissent dans peu de documents de la collection sont plus représentatifs du contenu de ces documents que ceux qui apparaissent dans tous les documents de la collection. Cette mesure est exprimée selon l'une des déclinaisons suivantes :
    - $\boldsymbol{Idf} = \log(\frac{n}{df})$
    - $\boldsymbol{Idf} = \log(N - \frac{df}{df})$

Lors des compagnes d'évaluation internationales, la mesure a eu des performances très limitées dans des corpus de taille très variable. Le problème posé est que les termes appartenant aux documents longs apparaissent très fréquemment et emportent le poids sur les termes appartenant à des documents moins longs. Les documents longs auront alors plus de chance d'être sélectionnés [**3**].

1.5. Requêtage

La recherche vise à sélectionner les documents pertinents qui couvrent les besoins d'information de l'utilisateur. Cette phase dépend de la représentation du document, les besoins d'information de l'utilisateur et les préférences de l'utilisateur (par exemple, la langue, la date, le format, etc.). Cette étape s'intéresse à l'expression des besoins modèles de recherche d'information de l'utilisateur, souvent à travers une liste de mots-clés représentant la requête.

Ainsi, la requête soumise par l'utilisateur subit les mêmes traitements que ceux réalisé précédemment sur les documents au cours de leur indexation. Toutefois, la requête peut être étendue ou reformuler pour renforcer les préférences des utilisateurs et le retour de pertinence. A la fin du processus de recherche, une liste de documents sera retournée [**1**].

1.6. Appariement

Les SRI intègrent un processus de recherche/décision qui permet de sélectionner l'information jugée pertinente pour l'utilisateur. A cet effet, une mesure de similarité (correspondance) entre la requête indexée et les descripteurs des documents de la collection est calculée. Seuls les documents dont la similarité dépasse un seuil prédéfini sont sélectionnés par le SRI.

La fonction de correspondance est un élément clé d'un SRI, car la qualité des résultats dépend de l'aptitude du système à calculer une pertinence des documents la plus proche possible du jugement de pertinence de l'utilisateur. [**1**]

Il existe deux types d'appariement :

1.6.1. Appariement exact

Le résultat est une liste de documents respectant exactement la requête spécifiée avec des critères précis. Les documents retournés ne sont pas triés.

1.6.2. Appariement approché

Le résultat est une liste de documents censés être pertinents pour la requête. Les documents retournés sont triés selon un ordre de mesure. Cet ordre reflète le degré de pertinence document/requête.

2. Les modèles de RI

Un modèle de RI [**2**] fournit une interprétation de la notion de pertinence. Il existe plusieurs modèle de RI textuelle développés dans la littérature. Le modèle joue un rôle central dans la RI c'est lui qui détermine le comportement clé d'un SRI. Il permet de donner une interprétation des termes choisis pour représenter le contenu d'un document. Ces modèles ont en commun le vocabulaire d'indexation basé sur le formalisme mot clé, et différent principalement par le modèle d'appariement requête-document.

On détermine le vocabulaire d'indexation par :

$$V = \{ti\}, i \in \{1 \dots n\}$$

Tel que n : mot ou racines de mots qui apparaissent dans le document

On définit le modèle de RI par un quadruplet (D, Q, F, R (q, d)) où :

D : ensemble de documents

Q : ensemble de requêtes

F : schéma du modèle théorique de représentation des documents et des requêtes

R (q, d) : la fonction de pertinence du document 'd' à la requête 'q'.

On distingue 3 principales catégories de modèles : modèles booléens, modèles vectoriels, modèle probabilistes

2.1. Modèle booléen

C'est le 1er modèle de RI, [**1**] il est basé sur la théorie des ensembles, dans ce modèle les documents et les requêtes sont représentés par des

ensembles des mots clés (termes) ou encore un vecteur booléen chaque document 'd' est représenté par une conjonction logique des termes non pondérés qui constitue l'index du document,

$$\text{Ex : } "d = t1 \wedge t2 \wedge \ldots \wedge tn"$$

La requête 'q' de l'utilisateur est représenté par une expression logique les termes relient par des opérateurs logique (AND (∧), OR (v) et NOT (ℸ))

$$\text{Ex : } "q = (t1 \wedge t3) \vee 1 \ (t2 \wedge t3))"$$

L'appariement RSV entre requête et document est un appariement exact, autrement dit, la fonction de correspondance est basée sur l'hypothèse de présence/absence des termes de la requête dans le document et vérifie si l'index de chaque document 'd' implique l'expression logique de la requête 'q'. Elle est définie comme suit :

$$\text{RSV} = \begin{cases} 1 & si\ dappartient l'ensemble d'écrit\ parq \\ 0 & sinon \end{cases}$$

2.2. Modèle vectoriel

C'est un modèle algébrique [5] où l'on peut présenter les documents et les requêtes par des vecteurs dans un espace multidimensionnel dont les dimensions sont les termes issus de l'indexation.

La comparaison de la requête au document est effectuée en comparant leurs vecteurs respectifs. On ramène aussi une proximité sémantique à une mesure de distance géométrique.

Formellement, si on a un espace 't' de Termes d'indexation de dimensions 'n', T={t1, t2, …, tn}, l'index d'un document 'd'est le vecteur $\rightarrow = (w_{1,d}, w_{2,d}, \ldots, w_{n,d})$, et une requête est également représenté par un vecteur $\rightarrow = (w_{1,q}, w_{2,q}, \ldots, w_{n,q})$ où :

Wi,d et wi,q correspondent aux poids du Terme 'ti' dans le document 'di' (respectivement dans la requête 'qi')et 'n' correspond au nombre de Termes de l'espace.

La fonction de correspondance mesure la similarité entre le vecteur requête et les vecteurs documents.

Une mesure classique utilisée dans le modèle vectoriel est le cosinus de l'angle formé par les 2 vecteurs, comme suit :

$$RSV(q,d) = \cos qd$$

Remarque :

Plus les vecteurs sont similaires, plus l'angle est petit et plus le cosinus de cet angle est grand. A l'inverse du modèle booléen, la fonction de correspondance évalue une correspondance partielle entre un document et une requête ce qui permet de retrouver des documents qui ne reflètent pas la requête qu'approximativement. Donc les résultats peuvent être ordonnés par ordre de pertinence décroissante.

2.3. Modèle probabiliste

Plusieurs approches comme [**6**] [**7**] [**8**] [**9**] [**10**] ont tenté de définir la pondération de façon plus formelle s'appuyant souvent sur la théorie des probabilités.

Le principe de base consiste à retrouver des documents qui ont en même temps une forte probabilité d'être pertinents, et une faible probabilité d'être non pertinents, étant donnée une requête d'utilisateur 'Q' et un document 'D', il s'agit de calculer la probabilité de pertinence du document pour cette requête.

Deux possibilités se présentent : 'D' est pertinent pour 'q', et 'D' n'est pas pertinent pour 'q'. les documents et les requêtes sont représentés par des vecteurs booléens dans un espace à 'n' dimension.

Exemple :

Soit un document 'dj' et une requête 'q' :

dj=(w1,j, w2,j, …,wn,j)

et q=(w1,q, w2,q, …,wn,q)

tel que : wk,j et wk,qϵ [0 1]

La valeur de $w_{k,j}$ (resp. $w_{k,q}$) indique si le terme 't_k' apparaît dans le document dj(resp. q) ou non. Le modèle de probabilité évalue la pertinence du document 'dj' pour la requête 'q'.

Un document est sélectionné si la probabilité que le document 'd' soit pertinent, notée p(R/D), est supérieure à la probabilité que 'd' soit non pertinent pour 'q'où 'R' est l'événement de pertinence et 'D' l'événement de non pertinence.

On note le score d'appariement entre le document 'd' et la requête 'q' :

$$RSV\left(\frac{Q}{D}\right) = \sum_{t \epsilon q \cap d} Log\ \frac{N}{Df_t}$$

Ces probabilités sont estimées par des probabilités conditionnelles selon qu'un terme de la requête est présent dans un document pertinent ou non. Par différentes formules on peut calculer la mesure de similarité (entre requête et document). Ce modèle a donné lieu à de nombreuses extensions, il est à l'origine du système OKAPI. Le modèle OKAPI BM25 a été développé par 'Robertson en 1994' dans lequel le calcul du poids d'un terme dans un document intègre des aspects relatifs à la fréquence locale des termes.

	Avantages	Inconvénients
Modèle booléen	• Le modèle est simple et transparent à comprendre pour l'utilisateur : ✓ Pas de paramètres	• La sélection d'un document est basée sur une décision binaire et Pas d'ordre

	cachés ✓ Il répond à une formule logique • Adapté pour les spécialistes et les vocabulaires contraints	pour les documents sélectionnés. • Pour beaucoup d'utilisateurs, la formulation de la requête n'est pas toujours évidente, elle est difficile. • Les tests effectués sur des collections d'évaluation standards de RI ont montré que les systèmes booléens sont d'une efficacité de recherche inférieure.
Modèle vectoriel	• La pondération améliore les résultats de recherche • La mesure de similarité permet d'ordonner les documents selon leur pertinence vis-à-vis de la requête	• La représentation vectorielle suppose l'indépendance entre termes
Modèle probabiliste	• La fonction d'appariement permet de trier les documents • Apprentissage du besoin d'information	• Problème des probabilités initiales et pas de langage de requête • Le modèle

		considère que tous les termes sont indépendants.

Tableau 1 : Les avantages et les inconvénients du modèle de RI

3. Les outils de recherche d'information

Il existe de nombreux outils de recherche d'information sur le Web, ces outils se spécialisent en fonction des services utilisés et du type d'information qu'ils recensent. On qualifie d'ailleurs très souvent aujourd'hui toute interface de recherche et d'interrogation de moteur de recherche et ce, quelle que soit la source interrogée et le système informatique utilisés [**11**]. Il convient en effet de distinguer les différents types d'outils de recherche sur l'Internet.

Dans le cadre de notre mémoire, nous distinguons trois catégories d'outils pour la recherche d'information sur le web: les moteurs de recherche, les annuaires et les méta-moteurs. Cette distinction qui repose également sur le mode d'indexation reste essentielle, car elle induit des usages et des technologies très différentes. Ainsi un annuaire thématique va-t-il référencer des sites Web, là où un moteur indexera toutes les pages d'un site ? En effet, l'annuaire facilitera le défrichage, le premier repérage des ressources dans un domaine ou un secteur défini par l'organisation arborescente proposée, alors qu'un moteur de recherche permettra de trouver un document très précis. Enfin les méta-moteurs permettent d'interroger en une seule fois différents outils de recherche, qu'ils soient de type annuaire ou de type moteur. Nous présentons dans ce qui suit ces trois catégorie d'outils pour la RI, et nous mettons l'accent sur les moteurs de recherche du fait qu'ils seront utilisés comme support de validation dans la partie contribution de notre mémoire.

Un moteur de recherche est une application permettant de retrouver des ressources (pages web, images, vidéo, fichiers, etc.) associées à des mots exprimant un besoin en information. Certains sites Web offrent un moteur de recherche comme principale fonctionnalité ; on appelle alors « Moteur de

recherche » le site lui-même (Google Vidéo par exemple est un moteur de recherche vidéo).

Ces outils de recherche sur le web sont constitués de « robots », encore appelés bots, spiders, crawlers ou agents qui parcourent les sites à intervalles réguliers et de façon automatique (sans intervention humaine, ce qui les distingue des annuaires) pour découvrir de nouvelles adresses (URL). Ils suivent les liens hypertextes (qui relient les pages les unes aux autres) rencontrés sur chaque page atteinte. Chaque page identifiée est alors indexée dans une base de données, accessible ensuite par les internautes à partir de mots-clés.

3.1. Les moteur de recherche

Les moteurs de recherche ne s'appliquent pas qu'à Internet : certains moteurs sont des logiciels installés sur un ordinateur personnel. Ce sont des moteurs dits desktop qui combinent la recherche parmi les fichiers stockés sur le PC et la recherche parmi les sites Web — on peut citer par exemple Exalead Desktop[1], Google Desktop et Copernic Desktop Search[2], etc.

Des modules complémentaires sont souvent utilisés en association avec les trois briques de bases du moteur de recherche. Les plus connus sont les suivants :

1. *Le correcteur orthographique :* il permet de corriger les erreurs introduites dans les mots de la requête, et s'assurer que la pertinence d'un mot sera bien prise en compte sous sa forme canonique.
2. *Le lemmatiseur :* il permet de réduire les mots recherchés à leur lemme et ainsi d'étendre leur portée de recherche.
3. *L'anti dictionnaire :* utilisé pour supprimer à la fois dans l'index et dans les requêtes tous les mots "vides" (tels que "de", "le", "la") qui sont non discriminants et perturbent le score de recherche en introduisant du bruit.

[1]https://exalead-desktop.fr.uptodown.com/windows
[2]https://www.copernic.com/fr/products/desktop-search/

En ce qui concerne les caractéristiques, les moteurs de recherche ont un fonctionnement commun, mais diffèrent par un certain nombre de critères [**12**]. Pour ce qui est du commun, rappelons simplement qu'ils procèdent tous avec les mêmes étapes :

- D'abord l'exploration du web, durant laquelle ils vont collecter les informations sur chaque page rencontrée.
- Puis l'indexation, durant laquelle ils vont enregistrer dans une base de données les informations collectées.
- Enfin la recherche, durant laquelle ils vont rechercher les données collectées en fonction des mots. Si tous les moteurs passent par ces étapes communes, ils ont tous leurs différences. Voici quelques éléments sur lesquels ils se différencient :
- D'abord la manière d'explorer le web.
- Ensuite, le choix des informations qu'ils vont récupérer des sites visités. Certains moteurs vont conserver le titre de la page, la description qu'en a fait le créateur de la page (metatag), parfois une partie du contenu de la page, …
- La manière de construire l'index, et sa taille. C'est d'ailleurs un des critères de performance le plus souvent mis en avant. En effet, plus un moteur indexe de pages, plus il a de chance de fournir un résultat correspondant à une requête. Google[3] passe pour avoir l'index le plus important. Les chiffres sont néanmoins difficiles à obtenir. En septembre 2005, Google indiquait avoir indexé 24 milliards de pages. Actuellement, pas plus de dix ans plus tard, ce nombre atteint 100 trillion de pages [**2**].

3.2. Les annuaires

L'annuaire (ou directory en anglais) est une liste de liens subdivisés en catégories suivant une structure en arbre, accompagnée d'une brève description. Bien que ce procédé fût pionnier en la matière, il tend à disparaître. En effet, le fait de devoir sélectionner les catégories dans lesquelles on recherche suppose que l'on sache exactement où chercher. Et on peut se demander où se positionne le site qui appartient à plusieurs

[3] http://www.google.com

catégories. Néanmoins, on doit reconnaître aux annuaires un gros avantage, celui de mettre en quelque sorte dans le contexte, ainsi les recherches dans la base de données sont diminuées, en plus d'obtenir des résultats plus pertinents.

Les annuaires sont donc des outils basés sur le recensement humain de l'information. Ils signalent des sites et des ressources de l'Internet comme un catalogue de bibliothèque signale des livres ou bien encore comme les pages jaunes signalent des entreprises. On distingue dans ce contexte deux catégories d'annuaires [**11**].

3.2.1. Les annuaires commerciaux

Ils se financent grâce à la publicité. Ils ont en principe une couverture dite "générale" (ils couvrent toutes les disciplines). Ils peuvent concerner le monde ou une zone régionale, citons :

- *Annuaires généralistes internationaux :* le plus connu est sans doute 'Yahoo Directory'[4], mais il existe aussi 'DMOZ'[5] de l'Open Directory Project et l'annuaire de 'Lycos'[6].
- *Annuaires régionaux commerciaux :* ce sont les annuaires qui recensent des sites en fonction de leur langue. Dans le cas des annuaires francophones, nous citons la version française de 'Yahoo Directory' ou encore l'annuaire 'Francité'[7].
- *Les annuaires qui recensent d'autre pays ou parties du monde:* comme l'annuaire 'Wohaa' pour l'Afrique et l'annuaire russe 'Yandex'[8] en Russie.

3.2.2. Les annuaires non commerciaux

Sont des annuaires élaborés par des individus de façon bénévole ou bien par des institutions. Ils sont soit généraux soit spécialisés. Leur préoccupation consiste toujours à identifier les ressources et les sites en tenant comptes de leur qualité. On distingue :

[4]http://yahoodirectory.net/
[5]https://dmoz-odp.org/
[6] https://search.lycos.fr/
[7]http://www.francite.com/
[8]https://yandex.com/

- *Annuaires à couverture (généraliste):* comme le 'Vlib'[9] (Virtual Library) et l'annuaire 'Resource Discovery network'[10].
- *Annuaires à couverture thématique ou spécialisée :* comme le répertoire en sciences humaines 'Voice of the Shuttle'[11] et le répertoire de ressources juridiques 'Findlaw'[12].

6.3. Les méta-moteurs

Ils sont des créations plus récentes. Ils constituent en fait la première génération des agents dits "intelligents". Ils permettent d'interroger en une seule fois différents outils de recherche, qu'ils soient de type annuaire ou de type moteur, afin de fournir une réponse plus exhaustive.

Il existe deux catégories de méta-moteurs: ceux en ligne et ceux consistant en un "logiciel client" à installer sur son ordinateur (le plus connu: COPERNIC) [**11**]. Le principe de fonctionnement des méta-moteurs est différent, certains indexent l'information contenue dans différents annuaires et moteurs, d'autres les interrogent simultanément de façon dynamique.

Certains de ces méta-moteurs retraitent plus au moins les réponses (tri, dédoublonnage). Ils permettent ainsi de rechercher de façon plus large sur le Web. Toutefois, cela peut également générer du "bruit" (réponses non pertinentes).

La parade mise en œuvre par certains méta-moteurs consiste à limiter le nombre de réponses de chaque outil interrogé (ce qui est indispensable et permet ainsi d'obtenir les réponses en principe les plus pertinentes).

4. L'évaluation

4.1. Définition

L'évaluation des approches de RI est nécessaire pour mesurer leur efficacité, leur performance et pour pouvoir les comparer en étudiant l'impact des différents facteurs employés dans ces approches.

[9] http://www.vlib.org
[10] https://www.tandfonline.com/doi/pdf/10.1080/13614570009516959
[11] http://vos.ucsb.edu/
[12]

Un système de RI efficace doit répondre de façon satisfaisante aux besoins d'information de l'utilisateur en termes de qualité des résultats retournes, de rapidité du système ainsi que la facilite d'utilisation du système qui représentent les principaux facteurs à évaluer pour un système de RI [**13**]. Dans notre cas et de manière plus générale en RI, on s'intéresse particulièrement à : la capacité d'un système à sélectionner des documents pertinents que l'on nomme efficacité (*effectiveness*). Le mode d'évaluation généralement utilise de nos jours est basé sur celui développé dans le projet Cranfield [**14**] communément appelé le paradigme de Cranfield. Ce paradigme définit la méthodologie d'évaluation des systèmes de RI en se basant sur trois éléments: une collection de documents sur laquelle les recherches sont effectuées, un ensemble de requêtes de test (besoins des utilisateurs) et la liste des documents pertinents pour chacune des requêtes (jugements de pertinence). L'idée générale de ce paradigme est de créer un environnement unique afin de pouvoir comparer les systèmes équitablement.

Cet environnement est appelé la collection de test.

4.2. Collection de test

La collection ou corpus de test est un aspect fondamental qui constitue le contexte d'évaluation, c'est-à-dire les éléments qui vont servir à tester le processus de recherche d'information. Généralement, chaque collection de test est caractérisée par une collection de documents, une collection de requêtes, et des jugements de pertinence des documents par rapport à ces requêtes.

Dans une tâche de construction d'une collection de test, les jugements de pertinence constituent la tâche la plus complexe. Les jugements de pertinence indiquent pour chaque document du corpus s'il est pertinent, et parfois même son degré de pertinence, pour chaque requête. Afin de construire ces listes de jugements des documents pour toutes les requêtes, les utilisateurs (ou un groupe d'évaluateurs) doivent examiner le contenu de chaque document, et juger s'il est pertinent par rapport à la requête.

Dans les campagnes d'évaluation tels que TREC[13], les collections de documents contiennent plusieurs millions de documents, ce qui rend impossible le jugement exhaustif de pertinence. Donc, dans le cas de grandes collections, les jugements de pertinence sont construits en se basant sur la technique de *pooling*, effectuée à partir des 1000 premiers documents retrouvés par les systèmes participants à l'évaluation.

Dans les années1990, l'US National Institute of Standards and Technology (NIST) a recueilli une grande quantité de données à travers la compagne de recherche TREC

. Au total, cela a abouti a une collection de test contenant 1.89 millions de documents, principalement constituée d'articles de presse issus de l'agence de presse américaine *Newswire*, ces derniers sont accompagnés des jugements de pertinence pour 450 "taches de recherche" présentées sous forme de requêtes crées par des experts.

Depuis l'an 2000, Reuters a mis à disposition une large quantité de ressources adoptée pour la classification de texte, le "Reuters Corpus Volume 1". Il est composé de 810.000 articles d'actualité en langue anglaise 3. Par la suite, un second volume est apparu contenant des données en 13 langues (néerlandais, français, allemand, chinois, japonais, russe, portugais, espagnol, espagnol latino-américain, italien, danois, norvégien et suédois).

Pour faciliter la recherche sur les collections de données massives tels que les blogs, la collection Thomson Reuters TextResearch Collection (TRC2) a été réalisée, avec plus de 1.8 million de documents.

Les taches d'évaluation Cross-language ont été menées au sein de Conférence and Labs of the Evaluation Forum (CLEF), traitant principalement des langues européennes.

[13]https://trec.nist.gov/data/test_coll.html

La référence pour les langues d'Asie orientale et la recherche multilingue est la NII Test Collection for IR Systems (NTCIR), lancée par la société japonaise pour la promotion des sciences [**1**].

4.3. Mesures d'évaluation

Les mesures d'évaluation permettent d'estimer quantitativement l'efficacité d'un système de RI [**1**]. L'objectif principal est de quantifier, pour chaque requête la capacité du système à retourner des documents pertinents, à savoir les ensembles des documents pertinents et des documents retournés par le système. Les documents pertinents non retournés par le système représentent l'ensemble de documents *silence* tandis que les documents non-pertinents retournes par le système génèrent du *bruit*. Un bon système retourne le maximum de documents pertinents (*minimiser le silence*) sans augmenter le nombre de documents non pertinents retournés (*minimiser le bruit*).

4.3.1. Rappel et précision

- *Précision :* est une mesure d'exactitude, elle varie entre [0,1] et elle est calculée de la manière suivante :

$$Précision = \frac{|TP|}{|TP + FP|}$$

  Et il existe beaucoup d'autres métriques et mesures qui peuvent servir à évaluer la précision d'un SRI.

- *Rappel :* est une mesure de perfection, elle varie entre [0,1], elle est calculée de la manière suivante :

$$Rappel = \frac{|TP|}{|TP + FN|}$$

Tel que :

- Les correspondances correctes trouvées par un système sont appelées « the true positives (TP) » et sont calculées ainsi :

$$TP = S \cap H$$

- Les correspondances incorrectes trouvées par un système sont appelées « the false positives (FP) » et sont calculées ainsi :

$$FP = S - S \cap H$$

- Les correspondances correctes omises par un système sont appelées « the false négatives (FN) » et sont calculées ainsi :

$$FN = H - S \cap H$$

Tel que : S est le système
Et H est l'humain

Les valeurs obtenues par le calcul des correspondances entre les requêtes et les documents ne sont comparables par le biais de la précision et du rappel, en fait, *le rappel peutprendre des valeurs importantes aux dépens de la précision*, en retournant toutes les correspondances possibles. En même temps, *la précision peut prendre des valeurs importantes aux dépensdu rappel*, en retournant que les correspondances correctes cependant peu nombreuses.

C'est pour ces raisons qu'il est préférable de prendre en considération les deux mesures simultanément via une mesure qui combine le rappel et la précision telles que : la F-mesure qui se calcule de la manière suivante :

$$F - Mesure = \frac{2 * (\text{Rappel} * \text{Précision})}{\text{Rappel} + \text{Précision}}$$

La F-mesure est une mesure globale de la qualité des correspondances produites, elle varie entre [0,1], cette mesure alloue la même importance à la précision et au rappel.

Une autre mesure de la qualité de l'alignement et qui combine le rappel et la précision : l'Overall qui se calcule de la manière suivante :

$$OVERALL = \text{Rappel}(2 - \left(\frac{1}{\text{Précision}}\right))$$

L'overall peut prendre des valeurs négatives si le nombre de fausses correspondances (FP) trouvées par le système dépasse le nombre de correspondances correctes (TP) trouvées par le système.

Dans la plupart des cas l'overall est plus petit que le rappel et la précision, ce qui rend difficile d'atteindre un overall supérieur à 0.5.

Une mesure pour évaluer le pourcentage d'erreurs du système automatique est le Fallout qui se calcule de la manière suivante :

$$Fallout = \frac{\mathrm{FP}}{\mathrm{FP} + \mathrm{TP}}$$

5. Conclusion

Dans ce chapitre, nous avons présenté les principales notions et concepts de la recherche d'information, des systèmes de recherche d'information et ceux des outils de recherche sur le web.

A travers les différentes sections que nous avons présentées, nous concluons que l'efficacité et la qualité des mécanismes (indexation et appariement requête/documents) mis en œuvre durant le processus de recherche ont un impact direct et déterminant sur les performances d'un SRI, en particulier sur la qualité des réponses. Lors de l'appariement requête/documents, seuls les documents qui sont les plus proches sémantiquement du besoin de l'utilisateur sont sélectionnés. De ce fait, plus les termes d'indexation ne sont représentatifs du contenu sémantique des documents et de la requête, plus la pertinence des documents sélectionnés est améliorée.

Dans le cadre de ce travail, nous souhaitons apporter des contributions pour améliorer la recherche d'information en prenant en compte le contexte et la sémantique.

Chapitre II : Indexation Sémantique

1. Introduction

L'indexation est une étape primordiale dans tout processus de recherche d'information. Sa qualité dépend de sa capacité à mieux représenter l'information portée par le contenu d'un corpus textuel. Dans les systèmes de recherche d'information classiques, l'indexation d'un document (ou d'une requête donnée), est réalisée par l'ensemble des mots-clés qu'il contient. Cette représentation, appelée aussi indexation classique, ne prend pas en considération la sémantique des mots pour décrire avec précision la thématique abordée dans le document (ou respectivement dans la requête), ce qui implique des résultats non pertinents lors de la recherche [**15**]. C'est ainsi que les travaux récents en recherche d'information proposent l'indexation sémantique, en se basant sur la représentation des documents (et requêtes) par les sens des mots (ou concepts), plutôt que par les mots eux-mêmes. Ce chapitre s'articule en quatre parties: la première partie présente la définition de web sémantique et ses principaux composants. La deuxième partie décrit l'indexation sémantique et conceptuelle.la troisième partie présente la base de connaissances et enfin, la dernière partie un aperçu sur les approches d'indexation sémantique.

2. Web sémantique

Cette dernière décennie a connu une évolution considérable de la gigantesque toile qu'est le web, marquée par la croissance permanente des données et des ressources qui y sont exploitées, Ce qui le rend le premier outil pour la production, la publication, la diffusion et le partage de l'information. Cependant la répartition à travers le monde d'un tel réseau d'informations, la croissance accrue du nombre de publications et la liberté totale d'y accéder ont révélé plusieurs limites et inconvénients. En effet, le web ne dispose pas d'outils pour décrire et structurer ses ressources de manière satisfaisante afin de permettre un accès pertinent à l'information. Par exemple, les liens entre les pages web, bien que porteurs de sens pour les utilisateurs, n'ont aucune signification exploitable par les machines.

C'est pour pallier ces insuffisances que Tim Berners Lee a proposé dans [**16**]d'étendre le web vers un web où l'information possédera un sens bien défini permettant ainsi aux applications d'exploiter directement la sémantique des ressources et de coopérer avec l'utilisateur afin de lui faciliter ses tâches.

2.1. Définition

Le Web sémantique (plus techniquement appelé « le Web de données ») permet aux machines de 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 automatisés d'accéder plus intelligemment aux différentes sources de données contenues sur le Web et, et de cette manière, d'effectuer des tâches (recherche, apprentissage, etc.) plus précises pour les utilisateurs.

Le terme a été inventé par Tim Berners-Lee, Co-inventeur du Web et directeur du W3C, qui supervise l'élaboration des propositions de standards du Web sémantique [**1**].

2.2. Principales composantes du web sémantique

La plupart du temps, lorsque l'on prononce le terme de Web sémantique, on parle des différentes technologies qui se cachent derrière. Parmi les plus connues, on peut citer RDF (Ressource Description Framework) qui correspond à un modèle d'information, et les formats d'échanges de données en RDF pour communiquer entre différentes applications (RDF/XML, RDF/JSON, N3, Turtle, N-Triples et d'autres).

Dans le domaine du Web sémantique, la sémantique des données est décrite par des ontologies avec des langages prévus pour fournir une description formelle de concepts, termes ou relations d'un domaine quelconque. Ces langages sont RDFS (Ressource Description Framework Schéma) et OWL (Web OntologyLanguage). Il existe aussi des langages de description des données structurées dans du XHTML afin que des outils effectuent un traitement automatique de ces différentes données. Ces langages sont RDF et Microformat[14] et nouvellement, arrivé avec HTML 5, Microdata. Voici d'ailleurs un article qui vous introduit le langageRDFa et un autre sur les Microdata. Ensuite, pour finir avec la liste des technologies, il existe un langage de requête, au même titre que SQL pour les bases de données relationnelles,

[14]http://microformats.org/about

SPARQL, qui effectue des requêtes, mais sur des triplets RDF. Il en existe d'autres (RQL et RDQL), mais ils sont bien moins utilisés.

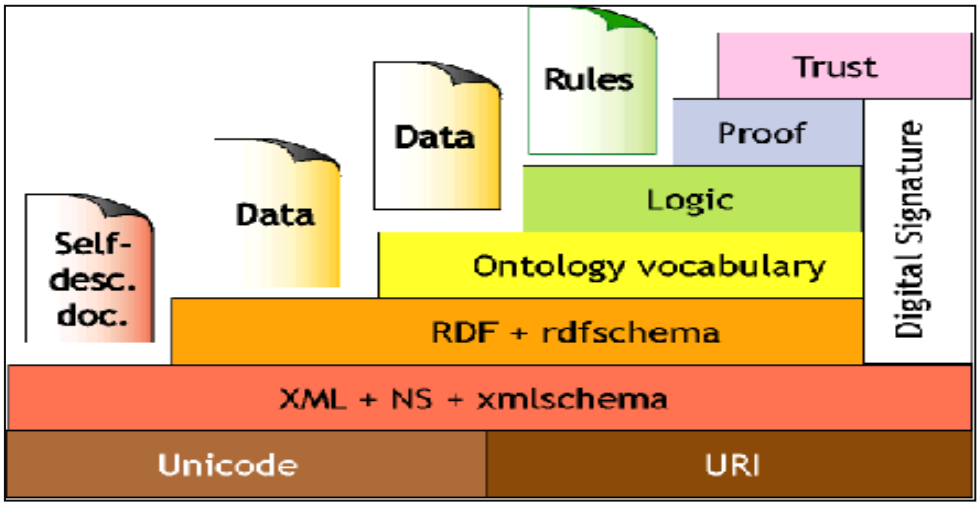

Figure 2 : Les couches du web sémantique.

XML (W3C, 1998) : fournit une surface syntaxique pour les documents structurés mais ne fournit aucune contrainte sémantique sur le sens de ces documents.

XML Schéma (W3C, 2001) : est un langage pour restreindre la structure des documents XML et étendre aussi XML avec des types de données.

RDF (W3C, 2004, c) : est un modèle de données pour les objets (« ressources ») et les relations entre eux, fournissant des sémantiques simples pour ce modèle de données qui peuvent être représentés en XML.

RDF Schéma : est un vocabulaire pour décrire les propriétés et les classes de ressources RDF.

OWL : ajoute plus de vocabulaire pour décrire les propriétés et les classes entre autres, les relations entre les classes, cardinalité, égalité, typage de propriétés plusriche, caractéristiques des propriétés et les hiérarchies des propriétés et des classes.

5.1. Indexation sémantique en recherche d'information

L'indexation sémantique ou l'indexation conceptuelle ont été présentées comme des alternatives pour pallier les défauts de l'indexation classique qui se base sur des

mots simples, dans la littérature plusieurs méthodes sont proposées dans chacune de ces 2 indexations [**17**]

3.1. Indexation sémantique

L'indexation sémantique s'intéresse principalement à la présentation des documents et requêtes par les sens des mots qu'ils contiennent plutôt que par les mots eux même. L'objectif sous-Jacent est d'améliorer la représentation des entités indexées et de pallier aux problèmes de l'indexation classique basée mots [**2**] .

Ce type d'indexation se base sur des algorithmes et les techniques de désambiguïsation sémantique (Word SenseDisambiguation WSD) pour indexer les documents et les requêtes avec le sens des mots simples. Une manière d'indexer serait par exemple, d'associer aux mots extraits, des mots du contexte qui aident à déterminer leur sens.Les travaux réalisés dans le cadre de l'utilisation du sens en RI ont rapporté des résultats mitigés.

Ona proposé une méthode de désambiguïsation basée sur le wordNet, qui est une structure conceptuelle organisée autour de la notion de synset.

Un synset regroupe des termes (simples ou composées) ayant un même sens dans un contexte donné. Ils sont liés par différentes relations telles que l'Hyperonymie (is-a) et son inverse l'Hyponymie (instance-de).

Gonzalo et ses collègues [**18**] ont réalisé des expérimentations sur l'indexation basée sens et l'indexation basée synset. Ils ont rapporté que les performances obtenues augmentent par rapport à l'approche basée sens. Pour Sanderson [**19**] et Krovetz et Croft [**20**] , une désambiguïsation « performante » permet d'améliorer les performances des SRI, notamment dans le cas des requêtes courtes [**17**].

3.2. Besoin de l'indexation sémantique

En indexation classique, les entités textuelles (requêtes et documents) sont représentées par des mots clés issus de leurs contenus, l'utilisation des mots pour représenter le contenu des documents et requêtes pose deux problèmes, l'ambiguïté des mots et leur disparité :

- **L'ambiguïté des mots :** aussi dite ambigüité lexicale, se rapporte à des mots lexicalement identiques et portant des sens différents. Elle est généralement divisée en 2 types [**21**]:

➢ L'ambiguïté syntaxique, se rapporte à des différences dans la catégorie syntaxique. Par exemple, « *play* » peut apparaitre en tant que nom ou verbe.

➢ L'ambiguïté sémantique, se rapporte à des différences dans la signification, et est décomposée en homonymie et polysémie selon que les sens sont liés ou non.

Le problème d'ambiguïté implique que des documents non pertinents, contenant les mêmes mots que la requête sont retrouvés.

- **La disparité des mots :** se réfère à des mots lexicalement différents mais portant un même sens. Ceci implique que des documents (pertinents) ne partagent pas de mots avec la requête, ne sont pas retrouvés.

Les travaux du domaine ont adressé ces problèmes séparément en apportant des solutions spécifiques à chacun d'eux, puis une solution globale s'est dégagée [**2**] .

4. Base de connaissances

La base de connaissances peut être considérée comme une base de données regroupant des connaissances spécifiques à un domaine spécialisé précis, sous une forme exploitable par une machine. Elle contient des règles, des faits ou d'autres représentations nécessaires à l'exercice d'une activité donnée pour laquelle cette base de connaissance a été développée.

4.1. Dictionnaire

Le dictionnaire est la base de connaissances la moins expressive. Il consiste en une collection de termes (un seul mot ou une séquence de mots) n'ayant aucun lien entre eux. Un cas particulier de dictionnaire est la folksonomie [**22**] dans laquelle les entrées sont collaborativement créées par une communauté.

Ces entrées, appelées tags, sont notamment utilisées pour indexer et organiser le contenu des pages web de manière collaborative [**23**]. Les Folksonomies feront l'objet d'une étude approfondie dans le chapitre 3 de ce mémoire.

4.2. Réseau sémantique

Quillian [**24**] défini le réseau sémantique comme une représentation permettant d'organiser les connaissances de la même manière que les êtres humains le ferait. Ainsi, un réseau sémantique peut être défini comme des concepts reliés à d'autres par des relations typées. Pour Medi [**25**], un concept (aussi appelé classe) représente une idée et inclut tout ce qui est caractéristiquement associé à elle.

En d'autres termes, c'est un élément symbolique représentant un ensemble de notions et d'idées. Par exemple, la Figure 3 représente le concept «Jus d'orange». Un «Jus d'orange» est composé de «orange» et d'«Eau». Dans notre exemple, « Lydia» peut être une «femme» qui a mangé une «orange» et bu du «Jus d'orange»…etc. [**23**] .

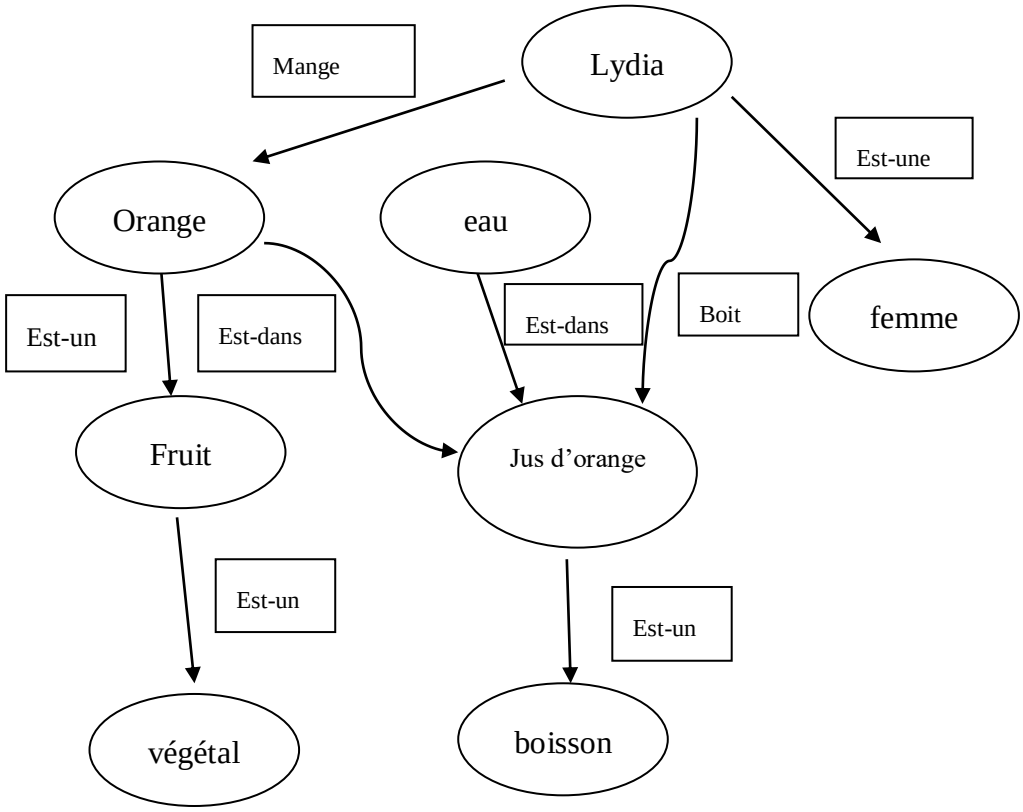

Figure 3: Réseau sémantique centré sur le concept "Jus d'orange"

4.3. Taxnomie

La taxinomie est un réseau sémantique où l'unique relation est une relation hiérarchique, transitive et non réflexive. Souvent, la relation hiérarchique «est_un» (en anglais,«is_a») est utilisée et on parle de classification. Les concepts sont

appelés des taxons. Un exemple classique de taxnomie est celui qui décrit les organismes vivants (Figure 4). L'avantage de l'inférence dans les taxinomies est mis à mal par le faible pouvoir d'expressivité de cette représentation [**23**] .

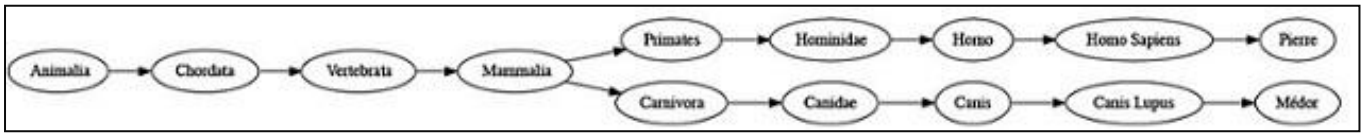

Figure 4 : Classification Biologique tirée de Wikipédia [**23**]

4.4. Thésaurus

Un thésaurus est une liste organisée de termes normalisés, validés, qu'ils soient descripteurs ou non-descripteurs (c'est-à-dire rejetés), reliés par des relations sémantiques (équivalence, hiérarchie, association, synonymie…) exprimées grâce à des signes conventionnels. Ces termes représentent les concepts d'un domaine de la connaissance et constituent un langage contrôlé pour l'indexation de documents et la recherche de ressources documentaires, selon des principes de construction élaborés depuis les années 1970 dans une norme internationale de l'ISO (dernière édition en 2011). L'indexation au moyen d'un thésaurus est fondée sur le principe combinatoire, la recherche se faisant par les opérateurs booléens (et, ou, sauf…) [**26**]. La Figure 5 montre les descripteurs pour le terme "Travail pénible" issu du thésaurus de l'Organisation Internationale du Travail[15] [**23**].

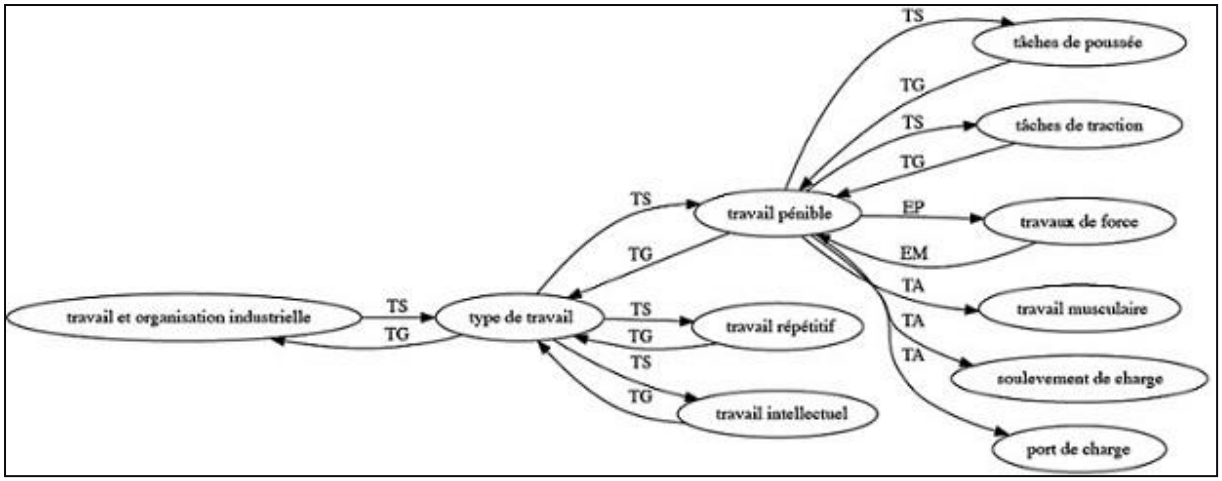

Figure 5: Extrait du thésaurus de l'Organisation Internationale du Travail [**23**]

[15] :http ://www.ilo.org/dyn/cisdoc/index_html ?p_lang=f

4.5. Graphe conceptuel et ontologie

Le graphe conceptuel est un formalisme introduit par Sowa en 1976 [**27**] [**28**] . A l'instar des réseaux sémantiques, il utilise une structure de graphe et introduit les notions de classes, de relations, d'individus et de quantifieurs. La force des graphes conceptuels réside dans leur expressivité car ils peuvent être traduits dans la logique des prédicats du premier ordre et donc permettre l'inférence.

L'ontologie, en informatique, est une représentation formelle de la connaissance introduite par Gruber en 1993 [**29**] : "spécification formelle et explicite d'une conceptualisation partagée".

Guarino&Giaretta [**30**] essaie de clarifier le doute existant sur cette définition et propose **7** définitions possibles. De par l'ambiguïté des définitions existantes, le terme "ontologie" est devenu passe-partout pour qualifier une base de connaissances possédant des relations sémantiques. Il existe, en revanche, des notions consensuellement acceptées par la communauté de l'ingénierie des connaissances :

- Concept (ou classe) et subsomption ;
- Relation entre concepts et subsomption des relations ;
- Individu (ou instance de classe) ;
- Relation entre individus ;
- Restriction ;
- Règle ;
- Axiome (ou fait).

De plus, il existe différents langages pour exprimer des ontologies, notamment l'Ontology Web Language (OWL) recommandé par le W3C [**23**].

4.6. Exemple de base de connaissance : wordNet

Une des bases de connaissances de référence dans le domaine de la RI est WordNet. Elle est le fruit d'un projet initié par Miller en 1985 [**31**] [**32**]. Abusivement, certains la considèrent comme une ontologie, mais en réalité c'est une base de données lexicale présentant des relations sémantiques. WordNetpeut être vue comme un réseau sémantique étendu. Il existe dans WordNet, plusieurs types de concepts : les termes (issus du lexique) et le sens des termes (appelé Synset). Par exemple, le terme «wood», peut désigner l'idée de matière (1), l'idée d'un regroupement d'arbres (2), l'idée d'une famille d'instruments de musique (3) ou encore l'idée d'un accessoire de golf (4).

Plutôt que de regrouper les concepts autour de leur forme lexicale, WordNet, à travers les synsets, regroupe les concepts en fonction de leur sens en contexte. Ainsi, WordNet construit deux types de relation :

1. Entre un synset et les termes employés pour le dénoter en contexte ;
2. Entre un synset et sa définition en contexte.

D'autres types de relation sont alors introduies entre synsets :

- Hyponymie : si X est hyponyme de Y alors X spécifie Y (le sens de «bois» est une spécification de la «végétation») ;
- Hyperonymie : si X est un hyperonyme de Y alors X généralise Y («végétation» est un hyperonyme du sens de «bois») ;
- Méronymie : si X est un méronyme de Y alors X est composé de Y (le sens de «bois» est un méronyme d'«arbre» dans le sens de la flore) ;
- Holonymie : si X est un holonyme Y alors X est une partie de Y (l' «arbre» dans le sens de la flore est un holonymede «bois»).

La Figure 6 [**26**] représente une infime partie de WordNet autour des sens du mot «wood», traduit en français. Le second sens du mot bois («Bois2») peut s'exprimer avec les mots «bois» ou «forêt» dans un contexte donné (ils sont synonymes et donc interchangeables en contexte) et «Bois2» est composé d'«arbres». La «Jungle» est un type de «Bois2» ; on peut alors déduire que la «Jungle» est composée d'arbres (cette relation n'est pas présente dans WordNet) [**23**].

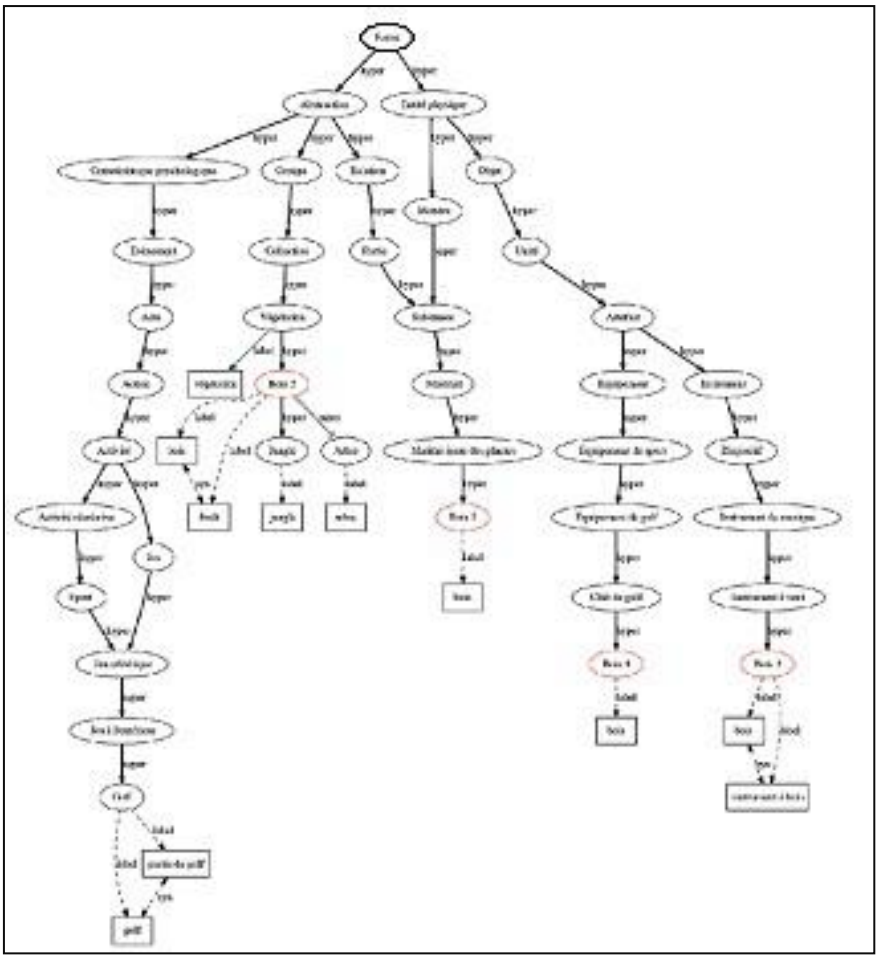

Figure 6 : Extrait traduit de WordNet [**23**].

5. Approche d'indexation sémantique

L'indexation sémantique a pour objectif la représentation des documents et requêtes par les sens des mots (ou les concepts) plutôt que par les mots d'indexation eux même. L'intérêt d'une telle approche est de lever l'ambiguïté des mots et de résoudre le problème de disparité des termes. Dans le reste de cette section, nous présentons une approche d'indexation sémantique de documents basée sur les CP-Nets.

Documents et requêtes sont alors indexés par des concepts. Les concepts sont extraits de WordNet, puis désambiguïsés. Des relations contextuelles sont ensuite découvertes entre concepts. Et enfin, concepts et relations associées sont organisées en un graphe CP-Net constituant l'index sémantique (ou conceptuel) du document. L'approche dans [1] s'articule autour des trois caractéristiques suivantes (résumés dans la figure 7 ci-dessous):

1. *une approche d'identification des concepts représentatifs du document*. L'approche est basée sur la projection des termes d'index sur l'ontologie WordNet et intègre une technique de désambiguïsation des concepts ambigus,
2. *une approche d'identification des relations entre concepts*. Cette approche est basée sur l'utilisation des règles d'association,
3. *une approche qui combine les concepts et les relations* correspondantes au sein d'un formalisme unifié, le CP-Net.

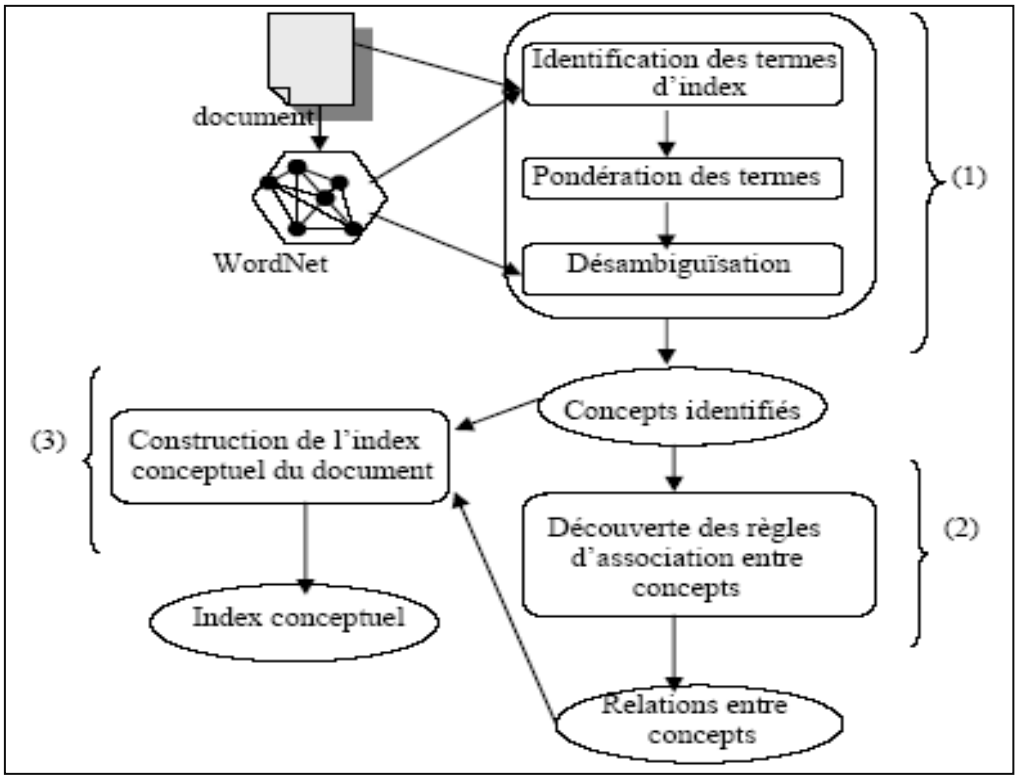

***Figure 7 : Les étapes de l'indexation conceptuelle basée CP-Nets* [1].**

6. Conclusion

Dans ce chapitre, nous avons exposé deux notions d'indexation (l'indexation syntaxique, et l'indexation sémantique), Nous avons conclu que la prise en compte des informations sémantiques dans le processus d'indexation peut améliorer la performance d'un système de recherche d'information.

Ces informations sont issues des ressources sémantiques telles que les ontologies et les thésaurus. Ces ressources sont de plus en plus disponibles. Tels que l'indexation consiste à extraire les descripteurs sémantiques à partir des documents. Ces descripteurs sont les concepts et les relations sémantiques entre ces concepts.

Chapitre III : Folksonomies

1. Introduction

Les mots-clés générés par l'utilisateur (les balises) ont été suggérés comme moyen léger d'améliorer des descriptions des ressources d'information en ligne et d'améliorer leur accès par le biais d'indexage. Le « marquage social » fait référence la pratique consistant à étiqueter ou classer publiquement les ressources. Dans un environnement en ligne partagé.

L'assemblage résultant de balises forme une « Folksonomie » : la « fusion » des mondes « Folk » et « Taxonomie » se référait à une notion informelle, organique assemblage de terminologie apparentée. Lorsque partagé avec d'autres, ou vues dans le contexte de ce que d'autres ont étiqueté, ces collections d'identifiants de ressources, les tags et les personnes commencent à acquérir une valeur supplémentaire par le biais des effets de réseau recherche de tags peut permettre la découverte de ressources pertinentes et des relations sociales qui se développent parmi les tagueurs deviennent un moyen de découverte de l'information en soi. Ceux qui utilisent activement des ressources Web, telles que des bases de données en ligne de les articles ont développés des outils permettant la création de collections personnelles de « signets » ou de pointeurs vers des ressources en réseau décrites ou « balisées » avec des mots identifiant et catégorisez-les. Ce document passe en revue les recherches sur le marquage social et la description socio-économique (identification environ 180 sources d'avant décembre 2007). Méthodes de recherche sur la contribution de l'étiquetage social et la folksonomie sont décrites, et les questions de recherche en suspens sont présentées. Un examen de la nature de la recherche en ligne, du contexte changeant pour recherche d'informations et efforts pour identifier d'autres sources d'indexation des ensembles de vocabulaire la scène pour cette discussion de la littérature émergente sur le social tagging et la folksonomy. C'est un nouveau domaine de recherche, où les perspectives théoriques et les méthodes de recherche pertinentes sont seulement maintenant être définies. Ce document fournit un cadre pour l'étude du marquage social et la folksonomie et l'analyse de leur contribution au paysage de l'information en ligne. [**33**]

2. Définition de folksonomie

La Folksonomie est le système dans lequel les utilisateurs appliquent des balises publiques aux éléments en ligne, généralement pour les aider à retrouver ces éléments [**34**]. Cela peut donner lieu à un système de classification basé sur ces balises et leurs fréquences, contrairement à une classification taxonomique spécifiée par les propriétaires du contenu lors de sa publication. Cette pratique est également appelée marquage collaboratif, classification sociale, indexation sociale et marquage social. Folksonomie était à l'origine "le résultat du marquage libre et personnel de l'information [...] pour sa propre récupération", mais le partage et l'interaction en ligne l'ont étendu à des formes collaboratives. L'étiquetage social est l'application d'étiquettes dans un environnement en ligne ouvert où les étiquettes des autres utilisateurs sont disponibles pour les autres. Le marquage collaboratif (également appelé marquage de groupe) est un marquage effectué par un groupe d'utilisateurs. Ce type de folksonomie est couramment utilisé dans des projets de coopération et de collaboration tels que la recherche, les référentiels de contenu et le bookmarking social.

Le terme a été inventé par Thomas Vander Wal en 2004 comme un valet de folk et de taxonomie. Les folksonomies sont devenues populaires dans le cadre d'applications logicielles sociales telles que le bookmarking social et l'annotation de photographies qui permettent aux utilisateurs de classer et de rechercher collectivement des informations via des balises partagées. Certains sites Web incluent des nuages de balises comme moyen de visualiser les balises dans une folksonomie.

Les Folksonomies peuvent être utilisées pour l'éducation de la maternelle à la 12e année, les affaires et l'enseignement supérieur. Plus spécifiquement, des folksonomies peuvent être mises en place pour les signets sociaux, les référentiels de ressources pour enseignants, les systèmes d'apprentissage en ligne, l'apprentissage collaboratif, la recherche collaborative et le développement professionnel.

Dérivée du "folk" et de la "taxonomie", la folksonomie est une caractéristique commune à la plupart des plates-formes de médias sociaux et Web 2.0. Il crée un

vocabulaire commun et partagé de termes ou mots-clés liés à chaque type de contenu unique.

La folksonomie survient principalement par le biais du marquage ou de l'ajout d'informations de métadonnées au contenu. Ces données peuvent être sous l'une des formes suivantes:

-Raccourci d'emplacement / URL / contenu

-Catégorie / type / classe

-Auteur / propriétaire

L'ajout de telles données permet d'améliorer la visibilité, la classification et la facilité de recherche du contenu. La folksonomie peut être large ou étroite.

La folksonomie large fournit une multitude de données de contenu et de balises, alors que les informations de folksonomie sont limitées.

2.1. Type de folksonomie

Une folksonomie apparaît lorsque les utilisateurs taguent du contenu ou des informations, tels que des pages Web, des photos, des vidéos, des podcasts, des tweets, des articles scientifiques et autres. Strohmaier et autres élaborer le concept: le terme "marquage" désigne une "activité volontaire d'utilisateurs qui annotent des ressources avec des" balises ", librement choisies dans un vocabulaire illimité et incontrôlé". D'autres expliquent les balises en tant qu'étiquettes textuelles non structurées ou mots-clés et qu'elles apparaissent sous la forme de simples métadonnées.

Folksonomies se compose de trois entités de base: utilisateurs, balises et ressources :

- Les utilisateurs créent des étiquettes pour marquer les ressources telles que: pages Web, photos, vidéos et podcasts.
- Ces balises permettent de gérer, classer et résumer le contenu en ligne. Ce système de marquage collaboratif utilise également ces balises pour indexer les informations, faciliter les recherches et naviguer dans les ressources.
- Folksonomie comprend également un ensemble d'URL permettant d'identifier les ressources auxquelles les utilisateurs de différents sites Web

ont fait référence. Ces systèmes incluent également des schémas de catégories permettant d'organiser des balises à différents niveaux de granularité.

Vander Wal identifie deux types de folklore: large et étroit. Une folksonomie étendue survient lorsque plusieurs utilisateurs peuvent appliquer la même étiquette à un élément, fournissant des informations sur les étiquettes les plus populaires. Une folksonomie étroite se produit lorsque des utilisateurs, généralement moins nombreux et incluant souvent le créateur de l'élément, marquent un élément avec des balises qui ne peuvent être appliquées qu'une seule fois. Bien que les deux saveurs, large et étroite, permettent de rechercher du contenu en ajoutant un mot ou une phrase associée à un objet, une vaste économie permet de trier en fonction de la popularité de chaque tag, ainsi que de suivre les tendances émergentes en matière d'utilisation des tags et de développement de vocabulaires.

Del.icio.us est un exemple de folksonomie large, un site Web sur lequel les utilisateurs peuvent baliser toute ressource en ligne qu'ils jugent pertinente avec leurs propres balises personnelles. Le site Web de partage de photos, Flickr, est un exemple souvent cité d'une folksonomie étroite. [**34**]

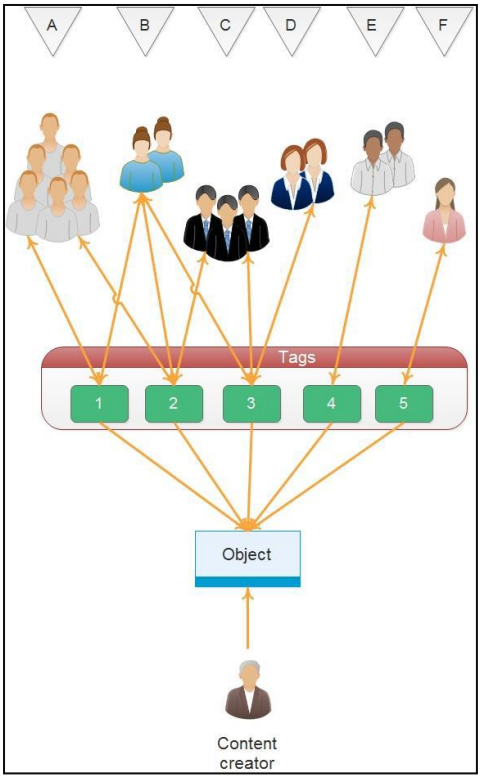

Figure 8: Folksonomy avec application à balises multiples («large») [**35**]

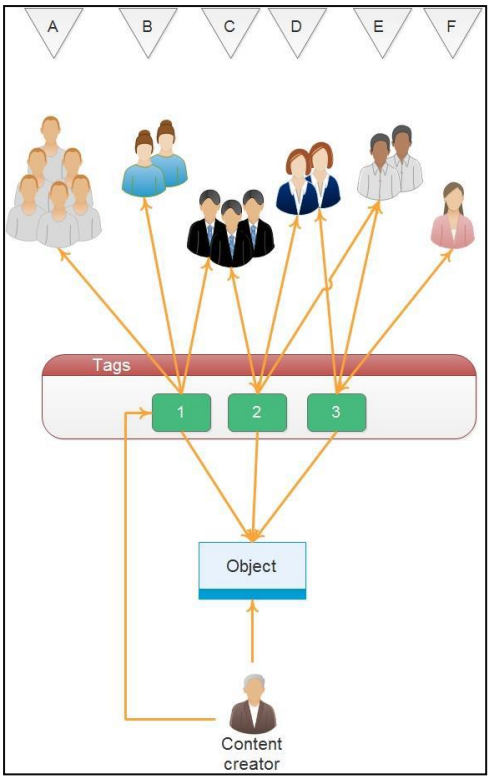

Figure 9: Folksonomy avec application à une seule étiquette ("étroit") **[35]**

2.2. Folksonomie Vs taxonomie formelle

Une comparaison entre folksonomie et taxonomie :

	Folksonomie	**Taxonomie**
Avantages	- Inclusivité des vocabulaires des utilisateurs de la communauté - Monnaie des descripteurs - Un dispositif peu coûteux à mettre en œuvre et à réutiliser - Plus de ressources navigables	- Précision accrue - Assisté professionnellement - Techniques élaborées de représentation des connaissances

Inconvénients	- Manque de précision dans la recherche d'informations - Pas de contrôle sur le vocabulaire - manque de hiérarchie	- Ensemble systématique de métadonnées (vocabulaire contrôlé) - Besoin de compétences d'expert en indexation

Tableau 2: comparaison entre folksonomie et taxonomie **[**34**]**

2.3. Caractéristiques de la folksonomie

-Rendre l'information accessible.

-Utilisation facile, multidirectionnelle et bidirectionnelle.

-Contribue à la formation du Web.

-Wiki, blogs dans le même mouvement.

-Tout le monde peut contribuer.

-Peut toucher tout le monde.

-Facilité de production tout le monde peut le produire facilement. Aucune compétence requise. Beaucoup d'outils sur le Web pour produire des signets.

-Peu coûteux.

-Permet à l'usager d'indexer des documents ou des informations et de les retrouver grâce à une classification des données à l'aide de mots-clés.

-Permettre aux internautes de mettre du contenu sur Internet et de partager leurs opinions.

-Permet d'accroître et de rendre plus rapide la diffusion de l'information.

-Les utilisateurs sont libres d'y choisir leurs propres mots-clés. La folksonomie est réellement centrée sur l'utilisateur. Il peut en faire une utilisation personnelle, professionnelle etc. De cette façon, l'utilisateur peut devenir plus structuré et cela lui permet d'organiser toutes ses informations.

-Permet l'ouverture d'esprit, c'est pourquoi la notion de partage est souvent évoquée.

Par exemple, il est possible de partager des photos ou des documents personnels que les autres usagers peuvent conserver et réutiliser au besoin.

Les consensus ne sont pas nécessaires en folksonomie. Il est alors impossible de lui reprocher les visions politiques ou encore les idéologies. [**36**]

3. Folksonomie et vocabulaire : étudier les tags

Les groupes intéressés par la recherche d'informations en ligne ont étudié la folksonomie à partir des perspectives de sa capacité à soutenir la recherche d'informations. Folksonomy est souvent critiqué parce que les étiquettes ne sont pas tirées d'un vocabulaire contrôlé. La terminologie agrégée dessinée de marquage devrait être intrinsèquement incohérent, et donc imparfait, selon théories de l'indexation. La littérature qui se concentre sur la folksonomie et l'information. [**37**]

3.1. Etudes générales

Merholz [**38**] célèbre le centrage de l'utilisateur sur le marquage. Sa métaphore, les chemins qui se portent dans l'herbe où les gens marchent plutôt que l'endroit où le paysagiste a voulu aller, est assimilé à «l'ethnoclassification» de Susan Leigh Star [**39**]. Mais il note le marquage inconvénients: synergie et inexactitude. Merholz réaffirme la valeur des connaissances acquises dans la classification et la récupération (2005), en soulignant la contribution des traditions de classification et contrôle du vocabulaire, ainsi que la valeur de la hiérarchie pour la désambiguïsation.

Quintarelli [**40**] définit la folksonomie comme une «*classification générée par l'utilisateur, émergeant par consensus ascendant* ». La vertu se trouve dans la

«*distribution ascendante distribuée et collaborative approche à la base* ", où" les *relations émergent naturellement* ". Le sens vient du l'agrégation de balises et leur regroupement. Aperçu de Quintarelli - axé sur la nature de vocabulaire - présente les inconvénients suivants de la recherche d'informations axée sur la folksonomie: «*Manque de précision*», manque de hiérarchie, faible quotient de découverte et problèmes de mise à l'échelle. Celles-ci contrastent avec les avantages, notamment le fait que les économies populaires "*reflètent les conceptuel de la population* "et sont capables de" *faire correspondre les besoins réels des utilisateurs langue*", en partie parce qu'elles sont inclusives. Le sentiment d'inévitabilité de Quintarelli, à savoir que les paysans sont un «*mouvement forcé*» (citant Shirky [**41**]), repose en partie sur le sentiment que les vocabulaires contrôlés sont «*peu étendus sur le plan pratique et économique*». Un résumé de 'side avantages sociaux' inclut les" *possibilités d'agrégation et d'analyse* "et la" *création de communautés* " dans le contexte d'une« *écologie des métadonnées* »émergente – la coexistence de vocabulaire contrôlé et folksonomie préconisé par Rosenfeld [**42**].

Kroski [**43**] décrit également les avantages et les inconvénients des stratégies descriptives folkloriques, fondées sur un examen de sites sociaux comme del.icio.us, 43 Things, Flickr et Technorati. Ses pros racontent tous à la «*sagesse des foules*» et inclure que les économies populaires "*sont inclusifs*", "*sont en cours*", "*Offre découverte*", "*sont non-binaires*", "*sont démocratiques et auto-modérateurs*", "*suis le désir lignes* "," *offrent un aperçu du comportement des utilisateurs* "," *engendrent une communauté* "," *offrent un faible coût alternative* "et " *offre la convivialité* ". Elle résume avec "*la résistance est futile*", citant le poids de information en ligne. Inconvénients: "*pas de contrôle des synonymes*", "*manque de précision*", "*manque de hiérarchie* "," *le problème du 'niveau de base'* "," *manque de rappel* ", et que - comme tout système social - les folksonomies "*sont susceptibles de jouer*".

Peterson [**44**] affirme que le relativisme de la folksonomie signifie qu'il est intrinsèquement imparfait, et échouera toujours à produire l'exactitude de la classification formelle (tirée de l'aristotelian catégories), au détriment de l'utilisateur. Weinberger, cependant, plaide pour la valeur du relativisme et la nécessité de tenir compte de l'inexactitude et de l'incohérence des connaissances systèmes de représentation [**45**]. Il voit le marquage de manière positive, pour sa capacité à intégrer de multiples perspectives, refléter les concepts émergents,

permettre l'émergence de groupes par la sémantique partagée, et tenir compte de la diversité [**46**], toutes les idées explorée plus tard dans son étude sur la longueur du livre [**47**].

3.2. Contrôle de vocabulaire et évolution

Les théories de la normalisation de soi affirment que les balises folksonomiques vont s'autoréguler - que le vocabulaire collectif deviendra plus cohérent avec le temps, sans imposition externe de contrôles. Dans une réaction aux contraintes du vocabulaire contrôlé, Shirky [**48**] affirme que le contrôle des synonymes réduit les nuances et sacrifie le sens, que la hiérarchie est souvent forcée et fausse et que la hiérarchie multiple est essentielle pour comprendre la nature multiforme du sens. Tout en admettant que «*le manque de précision est un problème*» avec la terminologie folklorique, il attribue cela au comportement de l'utilisateur, plutôt qu'à la nature même de la folksonomie, et prédit que les balises vont se normaliser d'elles-mêmes.

Les changements d'usage populaire sont reflétés dans les termes utilisés dans les systèmes de marquage. [**49**] Présente une analyse des balises del.icio.us pour "*Lois en puissance, Weblogs et L'inégalité* " [**50**] et illustre l'évolution du vocabulaire de marquage suite à l'introduction de l'expression 'longue queue'. Smith est sceptique quant au succès des balises populaires, citant le manque de contrôle sur la synonymie, la piètre précision des balises à mot unique et l'absence de hiérarchie [**51**].

Udell illustre la théorie de la stabilisation du vocabulaire autour de choix communs dans les balises dans une exploration interactive, montrant comment l'évolution du vocabulaire s'opère au sein du jeu de balises dans del.icio.us [**52**]. Cette stabilisation du vocabulaire a été confirmée par Millen et l'étude de Feinberg sur dogear chez IBM: "*les nouveaux tags diminuent progressivement en tant qu'utilisateurs finaux entrer plus de signets*", déterminé en calculant le pourcentage de nouvelles étiquettes en fonction de balises existantes et «*réutilisation considérable des balises*» [**53**].

Un certain nombre de questions doivent encore être explorées dans la relation entre Folksonomies et vocabulaires contrôlés. Il n'a pas encore été déterminé quelle sera la popularité et la les vocabulaires experts diffèrent et où chacun peut être utilisé au

mieux. Vocabulaires évolué avec le temps, et des études sont nécessaires pour déterminer si le marquage et la folksonomie pourraient fournir des indices sur l'évolution du vocabulaire. Par exemple, le marquage pourrait offrir un moyen simple de collecter des termes alternatifs à inclure dans un thésaurus existant. Les balises peuvent refléter l'émergence concepts qui présentent un intérêt pour l'entreprise et servent à capturer les connaissances des employés.

Le marquage pourrait permettre le développement d'une folksonomie d'entreprise pouvant être intégrée vocabulaire à l'échelle de l'entreprise (comme suggéré par Hayman, [**54**], 19).

3.3. Analyse de vocabulaire

Une folksonomie peut être étudiée en relation avec d'autres vocabulaires d'indexation pour identifier les contributions du marquage social. Lin et al. [**55**]décrivent les "*caractéristiques émergentes de la sociale classification*" à travers trois études de cas, de Connotea, Flickr, et del.icio.us, en regardant tag la distribution, les catégories de tags et la relation entre les tags et les termes d'index.

Kipp compare le vocabulaire des utilisateurs, des auteurs et des catalogueurs. Elle a d'abord analysé les tags, mots-clés d'auteurs et descripteurs fournis par des professionnels pour 176 entrées de citeulike.org. Les balises CiteUlike, les mots clés d'auteur et les termes fournis par l'indexeur ont été comparés sur une échelle de 7 points. Un nouveau vocabulaire est apparu, qui comprenait les «*balises de gestion du temps*» et les descripteurs (fournis uniquement par les indexeurs) (Kipp, [**56**], [**57**]). Dans une étude ultérieure, [**58**] ont examiné à nouveau les vocabulaires auteur, intermédiaire et de marquage (en s'appuyant sur Kipp, [**57**]). Cette fois, elle s'est concentrée sur des articles de biologie étiquetés dans CiteUlike et catalogués dans PubMed. Les résultats étaient similaires à ceux de l'étude précédente: la terminologie des tagueurs diffère de celle d'auteurs et indexeurs, bien que cela puisse être apparenté (sous forme de termes alternatifs). L'érudite communauté en utilisant l'outil de marquage CiteUlike a révélé un intérêt pour la méthodologie qui était non reflété dans le catalogage traditionnel; ces balises axées sur les méthodes ont fait une différence contribution à la description d'articles de biologie. Un résultat

similaire a été trouvé dans un CiteUlike étude portant sur l'information sur la santé [**59**].

[**60**]font état d'une étude examinant les étiquettes attribuées aux blogs dans relation avec les domaines, les parties du discours (à l'aide de l'outil POS de Stanford) et le sens (à l'aide de WordNet) pour voir ce qui peut être appris sur la sémantique des tags et le comportement des tagueurs individuels. La petite étude de Smith [**61**] explore la relation entre la folksonomie et analyse du sujet dans une étude des balises attribuées dans LibraryThing et les vedettes-matières attribué par le catalogue de la Bibliothèque du Congrès. Il est apparu, dans son petit échantillon, que Les balises LibraryThing étaient meilleures que les vedettes-matières pour identifier les sujets latents (celles qui ne respectait pas les règles de couverture pour le catalogage des objets dans les bibliothèques).

3.4. Marquer des espaces en tant qu'ontologies

Les développeurs de Web sémantique s'intéressent aux folksonomies en tant qu'ontologies émergentes. Mika [**62**] présente un modèle d'ontologies de balises «*Acteur-Concept-Instance*» et l'explore au sein de le contexte des réseaux sociaux. Dans ce modèle, les balises sont des "hypergères" qui représentes " *l'engagement d'un utilisateur* " qu'une ressource et un concept sont liés. Le contexte social est critique car il permet d'étudier le sens dans son "*émergence des actions des utilisateurs*" ...

"*Les ontologies sont nous: inséparables du contexte de la communauté dans laquelle elles sont créées et utilisé* ".

Ohmukai et al. [**63**] Décrivent la conception d'un système permettant aux utilisateurs de positionner leurs étiquettes. [Et signets] aux côtés de ceux des autres membres de leur réseau personnel, permettant à la construction d'une ontologie communautaire et inclure une proposition d'expression RDF du modèle. Beckett [**64**] propose une méthode d'utilisation de l'espace wiki pour désambiguïser et structurer balises en sémantique significative. Passant [**65**] propose aux utilisateurs un moyen de lier des tags à une ontologie exprimée en usage dans une sphère d'entreprise définie. Dans le système SemKey, Marchetti et Al. [**66**]Proposent une méthode pour faire la plupart

de ces tâches de manière automatisée, avec aide aux relations entre termes rendus explicites dans WordNet et Wikipedia.

Dans une étude plus théorique sur les ontologies émergentes, Dix et al. [**67**] Notent que«*Les systèmes de marquage collaboratif sont un mécanisme alternatif à l'approche du Web sémantique où les experts construisent des ontologies* » .Un «halo sémantique» de sens est construit autour d'un tag par interrogation automatique. Relations entre les balises, y compris plus large, plus étroit et termes connexes, sont calculés en tenant compte des changements temporels dans la définition. Ils proposent un algorithme qui prend en compte «l'agrégation» (toutes les étiquettes liées entre elles par occurrence), 'abstraction' (balises associées généralisées et spécialisées), 'ambiance' (contexte) et 'âge' (ambiance ordonnée dans le temps). Un test à petite échelle réalisé avec les données del.icio.us est signalé.

Halpin et al. En 2007 proposent d'extraire des ontologies significatives en se concentrant sur le «court la tête, plutôt que la longue queue, en examinant les points communs des étiquettes les plus utilisées, associées à mesures de la stabilité et de la valeur de l'information (mesure d'une étiquette en fonction du nombre de pages qu'il récupère). Leur modèle prend également en compte le cycle de rétroaction essentiel dans marquage. Schmitz en 2006 propose une méthode statistique pour induisant une structure sémantique à partir des tags de Flickr, qui repose également sur des similitudes, mais permet aux grappes de partager une certaine proportion d'un ensemble de caractéristiques (polythétique), mais pas nécessairement toutes les mêmes caractéristiques (monothétique). Il reste à déterminer si le marquage et la folksonomie permettent la découverte d'ontologie.
Les méthodes permettant d'exposer manuellement l'ontologie - ou de l'induire techniquement - doivent être étudié et raffiné.

3.5. Autres approches de la folksonomie en tant que vocabulaire

Il existe d'autres approches théoriques pour étudier le vocabulaire de marquage. Tennis se penche sur la folksonomie dans le contexte des théories de la classification. Plus précisément, il distingue classification de la bibliothèque (axée sur l'organisation par sujet et l'accessibilité physique), limites infrastructures (cadres qui relient des zones disparates d'une organisation (Bowker& Star et le social

tagging en 1999) (axé sur la gestion des informations personnelles), et explore les fonctions de chacun.En revanche, Voss affirme que le marquage représente simplement «Indexation manuelle», Tonkin place le tagging dans un contexte historique, en recherchant d'autres moyens de gérer les collections de documents et des approches contrastées basées sur des systèmes de fichiers locaux, à ceux basés sur Internet. Elle résume des recherches dans des domaines aussi divers que indexation, jeux de langage, communautés discursives, changement de code, herméneutique, indexation et les systèmes de mise en situation, d'annotation et de classement [et de classement], notant un continuum d'intérêts plutôt qu'une opposition. Voir «*annotations*» - au sens large, toute assertion ou métadonnée ressources numériques (y compris les tags) - en tant qu'opinion, nécessitant une connaissance de leur auteur pour évaluation. Sur la base d'études de cas (qui traitent des problèmes des systèmes existants), ils réclament une annotation pour des raisons économiques et de couverture, mais notent que «*savoir qui (ou quoi) à condition que les annotations puissent aider les gens à déterminer la pertinence des annotations pour leurs propres objectifs* » (Van Setten et al. en 2006). Berendt et Hanser en 2007 affirment que dans la blogosphère, les balises qu'un auteur attache à un article devraient être considérés comme «juste plus de contenu» parce que leur choix est une affirmation de la pertinence d'une poster sur un domaine d'intérêt particulier. Feinberg en 2006 examine la paternité et l'autorité – et en particulier le choix de la forme des tags - dans le contexte de différents objectifs de marquage (indexation des collections personnelles, partage de collections personnelles et création collective de collections fusionnées).

4. Folksonomies : un type d'indexation nouveau en complément des langages contrôlés

4.1. L'émergeant : les folksonomies

Les folksonomies peuvent apparaître comme un système mieux adapté au Web actuel. En tant que système d'indexation, elles disposent en effet de nombreux atouts:

-Mise à plat au lieu de structure hiérarchique qui permet de tout voir d'un coup au lieu de devoir entreprendre des recherches dans des catégories qu'on ne maîtrise pas.

-Navigation d'un nouveau genre permettant un affinement dans la recherche mais aussi des surprises.

-Multiplicité des usages: la typologie des relations est bien plus large et variée que la simple relation mots-matières - publications. Outre le fonctionnement à facettes (ancien, récemment réutilisé), les tags servent surtout des objectifs. La typologie (non exhaustive) établie par Joseph T. Tennis fait ressortir les buts suivants: identifier de quoi il est question, le type de ressources, organiser ses ressources, identifier les qualités et caractéristiques des ressources, organiser des tâches.

-Langage naturel plus proche de n'importe quel internaute.

-Coût cognitif de création plus faible que la catégorisation; dans le même sens, les auteurs de Le rôle... rapportent que «le classement à l'aide de tags n'est certainement pas vécu par les internautes comme un processus d'indexation». J. Dye indique que les tags sont plus adaptés à un cerveau qu'à une machine.

-Réintroduction de la vox populi dans la catégorisation: la classification des êtres vivants par exemple est née de la façon primitive dont tout un chacun a distingué les espèces, sur des critères simples (apparences, cris)... L'article Classification scientifique des espèces361 en donne un bon exemple: «devant l'inconnu, elle procède par extension et/ou assimilation: par exemple, la souris->la chauve-souris->le kiwi (couvert de poils, le kiwi était pour les Chinois assimilable à une souris végétale...)».

-Enrichissement du mouvement, à l'aide d'outils (avènement de Tags Clouds de deuxième génération, amélioration de la compréhension des besoins des usagers par les développeurs), d'apparition de bonnes pratiques et de minisyntaxes autour des tags, avec quelques exemples pour le site del.icio.us: mettre la source de l'information avec la syntaxe source_nom ou cite:source_nom; établir un lien de parenté avec une hiérarchie rudimentaire: parent_tag/sujet_tag; mentionner les noms des publications avec in:publication_nom; nommer le type de ressource: ressource_type; utiliser des synonymes ou des formes alternatives de tags. On constate aussi l'apparition de consensus que relève l'article Semioticdynamics and collaborative tagging366, voire même de tropes sur un site de description d'intrigues comme IMDb.

Les défauts des folksonomies viennent en partie de ses atouts: tout est au même niveau et peut devenir indigeste, causant de l'info-pollution; il y a des problèmes de

synonymie donc de bruit documentaire, et de l'imprécision (relevée par E. Kroski sur son blog Infotangle). Pour ces raisons, les folksonomies ont été baptisées par certains «Web sémantique du pauvre», en arguant qu'elles ne favorisent que la sérendipité. Elles connaissent pourtant un succès dont les raisons sont sa simplicité, la pertinence des termes agrégés à ceux des autres, l'assistance à l'indexation, des raisons psychologiques. [**68**]

4.2. Perspectives d'avenir des folksonomies au regard du Web sémantique

On aurait tort d'opposer complètement folksonomies et Web sémantique, comme l'attestent les travaux de T. Gruber ou N. Spivack. N. Spivack pense qu'on ne peut pas dire que les folksonomies soient meilleures que le Web sémantique, car elles sont à courte vue et ne représentent pas la même chose; elles peuvent donc se combiner. T.Gruber met en garde contre la confusion entre ontologies et taxonomies et plaide pour l'ontologie des folksonomies autour de 4 concepts (terme, document, taggeur et taggé).

Les discours sont relayés par des projets comme MOAT, et les microformats. Nul ne sait aujourd'hui à quoi ressemblera le futur Web 3.0 et quel y sera le rôle des folksonomies. [**68**]

4.3. Analyse comparée pour bibliothèques

Plusieurs analyses comparées ont été menées entre les folksonomies des sites d'indexation collaborative et les vedettes-matière, l'introduction des tags dans les bibliothèques étant trop récente pour qu'on puisse étudier ce type de folksonomie; on dispose néanmoins d'une étude danoise réalisée sur ceux de l'AADL., dans son article *Cataloging and You: Measuring the Efficacy of a Folksonomy for SubjectAnalysis*part du constat que les essais qui ont comparé langages contrôlés et folksonomies l'ont fait sous le seul angle de l'antagonisme, et prend Librarything comme point de départ; elle compare tags et LCSH pour 5 livres, en partant de l'hypothèse que les tags seraient meilleurs pour la fiction, et les LCSH pour les documentaires. Biais observés: les LCSH n'ont pas d'ordre de grosseur, les tags ne sont pas tous dans le Tag Cloud pour les livres très populaires. En conclusion, il est difficile de trancher, mais Librarything semble être le plus performant lorsqu'il n'y a pas trop de problèmes de synonymes; surtout, c'est un espace où elle a fait de

nombreuses découvertes. Louise F. Spiteri analyse des tags sur del.icio.us, Furl et Technorati à la lumière des recommandation du guide NISO, notant la présence de nombreux synonymes (22% sur del.icio.us), l'utilisation des tags pour désigner surtout des objets, l'usage courant des abréviations, mais assez peu des néologismes rejetés par le guide NISO. Les tags sont finalement plus conformes aux recommandations professionnelles qu'on pourrait le penser. Pour l'insertion des tags dans les bibliothèques, elle recommande de permettre la construction en plusieurs mots et de liens vers des dictionnaires en ligne. Johanna Grandström étudie 500 tags de l'AADL; elle se demande ce qu'ils désignent, quelles sont les différences entre fiction et documentaire et comment ils diffèrent de l'indexation professionnelle. Son étude montre que les tags désignent surtout le sujet de l'œuvre, que ceux assignés à la fiction sont plus multidimensionnels, et que les tags désignant expériences et tâches, employés dans le monde des folksonomies, ne le sont pas ici; les usagers n'usent pas des termes du catalogue, et les tags sont généralement moins spécifiques que les termes professionnels.

La communauté des mangas et anime semble vivace, «utilisant ses propres termes qui de bien des façons étaient différents des professionnels.» Ces études convergent vers le constat que la coexistence des deux est enrichissante. Les folksonomies sont aussi avantageuses pour l'accès à de l'information rapide, pour la veille, alors que l'indexation professionnelle permet de «*sélectionner de l'information à caractère plus durable*.» [**68**]

4.3.1 Pourquoi, comment fonctionnent les folksonomies en bibliothèque ?

En bibliothèque, traditionnellement, l'indexation se fait de manière professionnelle, et doit respecter des contraintes communes, des normes, pour pouvoir utiliser de la même façon le catalogue de plusieurs bibliothèques, sans que cela ne nécessite de formation particulière. Pour ce faire, le professionnel décrit précisément le contenu du document indexé. Cela est censé garantir la qualité de l'indexation, et permet d'aboutir à un résultat précis dans un large catalogue (Babelthèque, 2010). L'usager doit connaître les normes en vigueur pour pouvoir s'orienter, ou bien demander l'aide d'un professionnel pour sa recherche. Il doit donc être au sein de la bibliothèque, ou bien utiliser une messagerie instantanée pour dialoguer avec le bibliothécaire. Mais depuis la prise d'ampleur des technologies du Web, les

bibliothécaires voient leurs ressources augmenter, et l'indexation traditionnelle ne suffit plus pour tout indexer.

De plus, comme nous l'avons vu supra, les utilisateurs fréquentent moins la bibliothèque. C'est une perte pour la transmission du savoir, qui est la raison d'être des professionnels de l'information et de la documentation, et d'autre part, si les recherches entreprises par les utilisateurs n'aboutissent pas à un résultat satisfaisant, ils se tourneront vers d'autres sources de connaissance, qui sont possiblement non validées et fausses. Pour ne pas laisser les usagers dans l'ignorance, et rester dans la mouvance des nouvelles technologies, les bibliothécaires ont décidé de changer, de se former, et d'entrer dans une nouvelle ère en impliquant le public, avec notamment les folksonomies.

Leur fonctionnement est simple : l'usager n'a qu'à ajouter des mots-clés, de manière libre, sur les notices bibliographiques des livres qu'il souhaite. « *Les internautes peuvent ranger leurs livres dans des étagères virtuelles en ajoutant librement des* « tags » *ou des* « étiquettes » *de classement* ». Les usagers choisissent des mots-clés qui caractérisent le mieux l'œuvre selon eux. [**69**]

4.3.2. Les folksonomies et la bibliothèque

Les folksonomies et l'indexation sociale visent la reconquête de l'usager dans la bibliothèque, certes, mais pas uniquement un retour dans la bibliothèque physique. Elles ont aussi une visée plus large, qui est le développement de la bibliothèque numérique, qui est en pleine expansion. La bibliothèque numérique est un maillon de la chaîne des bibliothèques, il n'est plus possible de l'ignorer.

La bibliothèque physique est une institution culturelle, il n'est pas question de la remplacer, ni de la supprimer. Elle fait partie du patrimoine des villes, architecturalement, historiquement. Mais la bibliothèque numérique ne cherche pas à contrer la bibliothèque physique, au contraire, elle veut cohabiter avec elle. Les folksonomies, qui ramènent l'usager à une place centrale, lui donnent une nouvelle vision de la bibliothèque: celle d'un lieu moderne, présent dans les murs et hors les murs.

Nous pouvons faire un parallèle avec les musées virtuels. Est-ce parce qu'une personne a vu une exposition sur le site du musée qu'elle ne se déplacera pas pour voir réellement cette exposition ? Le monde virtuel donne aux usagers des clés pour appréhender la réalité, et les actions numériques peuvent aider l'insertion dans un

monde réel plus difficile d'accès, comme un musée. En effet, certaines œuvres sont plus appréciables lorsqu'on connaît leur histoire, et un musée bondé ne permet pas toujours de s'approcher des œuvres pour lire leur intitulé. Au contraire, sur Internet, un clic donne accès au nom de l'œuvre, à la biographie de l'artiste, à la date, à l'histoire de l'œuvre. Il en est de même pour la lecture. Si les usagers cernent bien le lieu dans lequel ils vont évoluer, ils le démystifient et peuvent l'aborder plus sereinement.

Des ponts doivent être créés entre le physique et le virtuel, et en cela les folksonomies agissent, faisant à la fois la promotion à la fois des bibliothèques traditionnelles, et des bibliothèques numériques. [**69**]

4.3.3. Les questions soulevées par les folksonomies

a. La question de l'autorité

Avant l'arrivée du web2.0, ce qui était publié sur le web provenait d'un auteur qui est le plus souvent connu du lecteur. Aujourd'hui, cette notion d'autorité est transférée du côté de l'usager, remplacée par son rôle actif. L'usager peut aussi bien être auteur que lecteur. Il apporte une valeur au contenu, aux ressources. Mais l'on perd alors la notion d'autorité, il faut donc faire un nouvel usage des informations trouvées, s'adapter en prenant en compte le rôle nouveau des utilisateurs.

Il ne faut pas confondre pertinence et popularité. En effet, un lien populaire sera plus souvent taggé, et parfois par des sites publicitaires, il faut donc être attentif à la pertinence plutôt qu'au nombre de mots-clés ajoutés. La solution est de former les internautes, il faut leur donner les connaissances nécessaires pour garder une validité dans l'indexation, et dans la recherche. Cette formation n'est pas évidente à mettre en place, de par sa nouveauté, et le bibliothécaire y joue un rôle actif. [**69**]

b. Quel avenir ?

Une solution au problème des folksonomies dans l'indexation en bibliothèque2.0 serait un modèle de bibliothèque hybride, c'est-à-dire un modèle dans lequel l'usager pourrait tagger les ressources, avec un contrôle de l'indexation réalisé par le bibliothécaire. La difficulté provient du nombre de ressources et le nombre de tags, ce qui nécessiterait un travail constant du bibliothécaire, qui ne pourrait alors pas se consacrer à d'autres activités.

Une autre forme de contrôle permettrait aux utilisateurs de tagger seulement leurs propres ressources, qu'ils maîtrisent mieux, avec l'utilisation de folksonomies étroites. En effet, si l'usager a fait l'effort de publier un document, ou d'ajouter du contenu à sa bibliothèque, cela

implique que ce document lui tient à cœur, et qu'il l'indexera en conséquence, pour que les autres utilisateurs puissent le retrouver lors d'une recherche. Mais comme nous l'avons vu supra, cela pourrait nuire à la richesse de l'indexation sociale.

Le contrôle peut aussi s'effectuer avec l'aide d'un dictionnaire qui refuserait le mot-clé s'il ne le reconnait pas. Cela reviendrait à normaliser les folksonomies, et l'on peut alors obliger l'utilisateur (et cette pratique est déjà répandue) à insérer des séparateurs, qui seraient une virgule, un point-virgule. Cette contrainte semble facile à mettre en place et ne réduit pas la richesse des folksonomies.

Il est plus difficile de contrôler la subjectivité des tags, cette vérification ne peut se faire de manière automatique, mais seulement par un professionnel. Cela réduit de beaucoup les possibilités, car au vu du grand nombre de ressources trouvées dans une bibliothèque numérique, virtuelle ou physique, ce serait un travail colossal.

Certaines bibliothèques, virtuelles ou numériques comme Babelio[16] permettent aux utilisateurs de participer non seulement à l'indexation, mais aussi au contrôle, faisant ainsi preuve d'une grande confiance en ses usagers. Si un mot-clé est jugé pas assez approprié ou peu pertinent, il est signalé aux modérateurs du site qui le suppriment

La solution serait de trouver un juste milieu entre un catalogue rigide associé à un thésaurus et une folksonomie débridée pour pérenniser l'usage des folksonomies et, plus généralement, de l'indexation sociale. Cette solution pourrait se matérialiser dans un contrôle lui-même collaboratif, fait mutuellement par les usagers comme c'est le cas sur le site Babelio. De plus, cela entre parfaitement dans les cadres du web2.0, qui est lui-même un web de partage, ainsi que dans l'idéologie des folksonomies. Cela permettrait d'écarter un contrôle automatique, trop rigoureux, trop rigide, qui ferait perdre une grande partie de la richesse de cette nouvelle forme d'indexation. [**69**]

c. Les folksonomies ramènent-elles réellement l'usager dans la bibliothèque ?

Deux mondes semblent être en parallèle [**69**] : d'une part les blogs, ou catalogues de bibliothèques physiques, qui peinent à trouver des taggeurs, comme Bib'n Blog[17] d'autre part la bibliothèque de Babelio, qui foisonne de tags, et qui laisse le contrôle à ses usagers très actifs. D'où vient cette différence ?

A cette question on pourrait répondre que Babelio est une seule bibliothèque pour tout le monde, à la différence de la bibliothèque d'Auch qui n'est connue que par les

[16] Voir questionnaire proposé aux fondateurs de Babelio
[17] Voir entretien avec Jean Gulli

habitants de cette ville. Pierre Fremaux, un des cofondateurs de Babelio, nous explique qu'en termes de chiffres, Babelio ne regroupe qu'une dizaine de personnes par ville. S'il existait un site par ville, la fréquentation serait très faible.

Cela explique donc les difficultés de Bib'n Blog, qui peine à trouver des usagers hors les murs pour participer au fonctionnement du blog.

Ce constat nous amène à penser que l'indexation sociale et les folksonomies, à elles seules, ne suffisent pas à reconquérir les usagers, car ceux qui sont actifs sur le site des bibliothèques sont aussi ceux qui fréquentent les bibliothèques physiques. Cet effort de modernisation n'a pas encore fait changer les usages de fréquentation et n'a pas amené de nouvelles personnes à la bibliothèque, physique, numérique, ou 2.0.

Cependant, nous pouvons penser que cette évolution se fera lentement, mais que les démarches entreprises finiront par être efficace, car l'image de la bibliothèque, qui sort de ses murs, qui est présente en ville, sur Internet, dans les pensées, va changer et perdre son aspect désuet.

Le bibliothécaire a une nouvelle mission : redorer l'image du monde des livres, sans trahir la mission des bibliothèques publiques, en évoluant, en restant à la page, en écoutant son public et en répondant à ses demandes. Les folksonomies sont une des voies de ce nouvel élan.

5. Applications utilisant les folksonomies (RI)

Un logiciel social permet aux individus de communiquer, de se connecter ou d'échanger internet. Il permet aussi de constituer une communauté ou un groupe virtuel.

La fonction principale du logiciel social consiste en une mise en contact.

Plusieurs outils sont utilisés dans les applications des réseaux sociaux par exemple :

-Instant messaging : cette application permet de communiqué avec une autre personne via une connexion.

-Internet relay chat (irc): permetà plusieurs utilisateurs de communiquer entre eux.

-Internet forums : permet aux utilisateurs de poster un sujet permettant aux autres de réagir ou de commenter

-Blogs : sont des journaux en ligne qu'une personne alimente régulièrement avec des billets « post en anglais »que d'autres peuvent commenter.

-Wikis : un ensemble de pages web qui peuvent être créées facilementgrâce a la syntaxe de wiki.

-Social network services : permet aux utilisateurs de se rencontrer en ligne et de partager une passion ou un but commun.

-Social guides : recommandent des lieux à visiter comme des cafés, restaurant, bornes wifi.

-Social bookmarking : un site qui permet de partager les sites favoris (bookmarks) des utilisateurs entre eux.

-Social citations : destiné avant tout au milieu académique, ce logiciel permet de poster des citations des articles se trouvant sur internet.

6. Les travaux connexes

5.1.1. Les travaux de Claire Lebreton

Claire Lebreton présente une analyse générale du contexte de l'indexation en bibliographique et dans le mode de l'indexation social (I) puis de détailler les expériences des bibliographiques qui pratiquent l'indexation social afin de dresser un bilan des enjeux, opportunités et risques de ces pratiques en bibliographique.

5.1.2. Les travaux de Imen Ben sassi

Imen Ben sassi propose un travail dans le cadre de l'exploitation dans un environnement mobile plus précisément le travail porte sur les 4 volets suivants :

a) Définition et modélisation du contexte
b) Exploitation de la situation dans le domaine de la RI
c) Exploitation des données contextuelles dans le cadre de la recommandation d'amis
d) Proposition d'un cadre d'évaluation pour la RI contextuelle

5.1.3. Les travaux de Chahrazed Bouhini

Elle propose un travail avec un objectif d'amélioration des résultats de RI classique. Les informations sociales (relations sociales, annotations, clics, profil… etc.) peuvent être exploitées au sein même du modèle de RI (modèle de document et de

requête, fonction de pondération/ de correspondance), et des exemples détaillés d'informations sociales utilisées pour améliorer les résultats de systèmes de RI.

	Domaine	**avantages**	**Inconvénients**
Imen Ben Sassi	Environnement mobile	-Fournis aux utilisateurs des éléments plus pertinents	-Obligation des nouvelles technologies
Claire Lebreton	Bibliothèque	-Mettre en garde contre la confusion entre l'inélégance collective et les pratiques collaboratifs	-Interaction difficile entre la bibliothèque et la folksonomie
Chahrazed Bouhini	Réseaux sociaux	-Adaptation des composants de modèle proposé dans un cadre social	-Pas de collection de test standard pour la RI

Tableau 3: tableau des travaux connexes sur la folksonomie

7. Conclusion

Dans ce chapitre, nous avons introduit les notions de marquage social et présenté notre contribution à une technique d'indexation basée sur la folksonomie. Les résultats obtenus sont très encourageants et nous incitent explorer l'utilisation des folksonomies comme un complément permettant de tirer profit du pouvoir toujours croissant des masses alimentées par l'utilisation d'internet.

Chapitre IV:
Conception et modélisation de la solution

1. Introduction

Différents systèmes d'organisation des connaissances (KOS) aident à prendre en charge une indexation sophistiquée des documents. Les exemples courants de KOS incluent les systèmes de classification (taxonomies), les thésaurus et les mots clés contrôlés.

Ces dernières années, le “tagging social” s’est imposé au sein du Web2.0 comme le principal moyen de classification de données massives. En effet, l’emploi pour la classification d’ensembles de tags non contrôlés, appelés folksonomiesaméliore les techniques de récupération et aide les utilisateurs à déterminer la pertinence d’un document. Au cours de la dernière décennie, un problème bien connu de l’indexation de documents avec les métadonnées de description de contenu a été traité dans une nouvelle perspective centrée sur l’utilisateur. Dans le cadre du «Web 2.0», les internautes ont commencé à publier leur propre contenu à grande échelle et ont commencé à utiliser un logiciel social pour stocker et partager des documents, tels que des photos, des vidéos ou des signets. En outre, ils ont commencé à indexer ces documents avec leurs propres mots-clés pour les rendre accessibles. Dans ce contexte, les mots-clés attribués sont appelés «balises», le processus d'indexation est appelé «marquage social» et la totalité des balises utilisées au sein d'une plate-forme est appelée folksonomie.

Dans ce chapitre, nous allons présenter la folksonomie comme un moyen rapide et personnalisé de représentation et d’indexation de l’information.Elles sont aujourd’hui considérées comme une alternative moderne aux ontologies et aux ressources lexicales traditionnellement utilisées dans la RI. En outre, nous nous proposons d’élaborer un algorithme de pré-indexation qui consiste à vérifier l’existence des mots clés d’un document qui vient d’être ajouté à la base de données dans la Folksonomie de tags existants.

2. Organigramme de la solution

La figure ci-dessous, illustre la logique de notre solution sous forme d'organigramme.

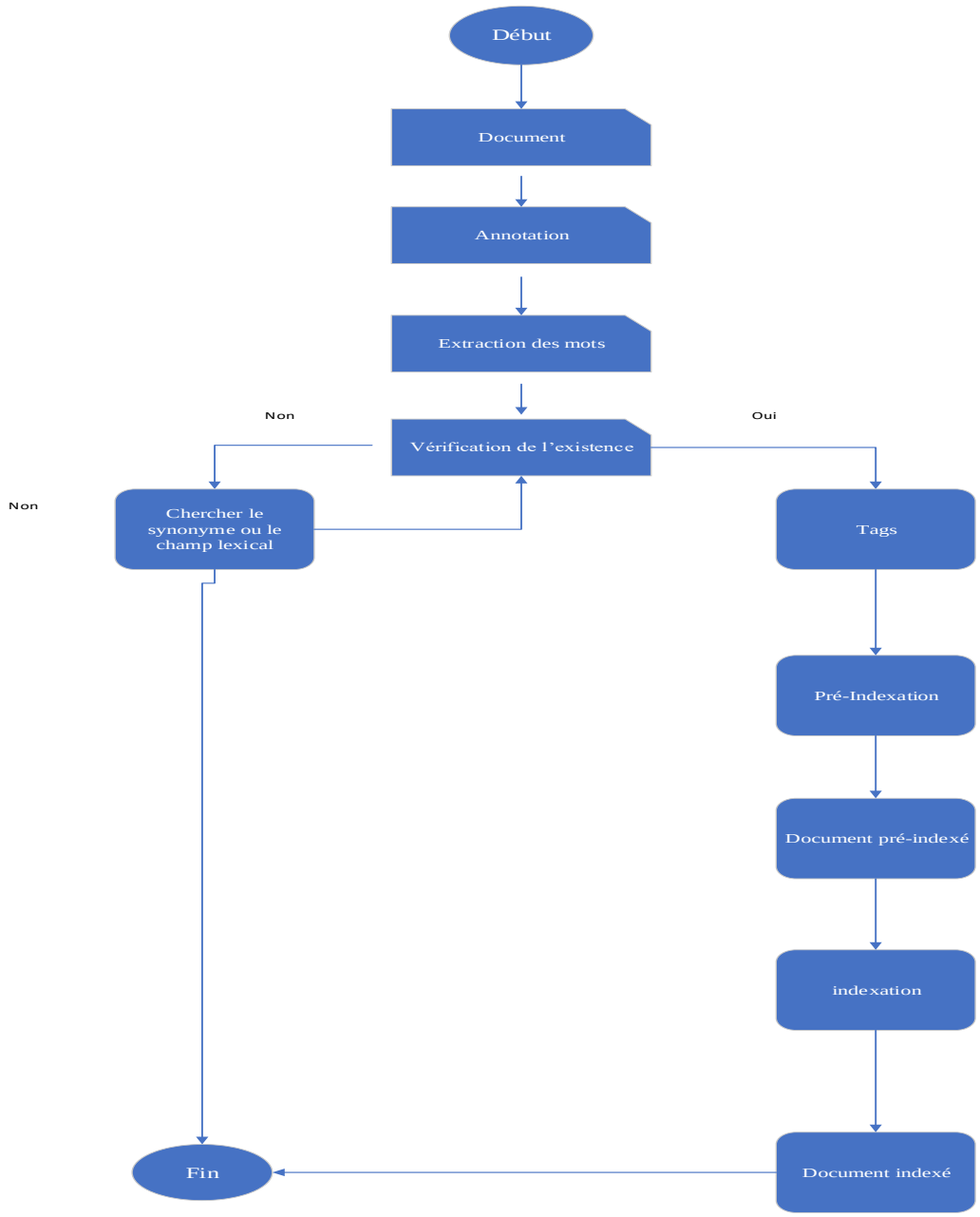

Figure 10: organigramme de pré-indexation

2.1. Description de l'organigramme

Le processus d'indexation est illustré par l'organigramme donné par la figure 10. Il débute par l'acquisition du document, Suivie par l'extraction des principales

informations du document ou ce que l'on appelle Annotation. Après cette dernière, nous entamons l'extraction des mots clés, ensuite nous passons à une autre phase qui commence par la localisation des mots clés dans la Folksonomie à l'aide d'une fonction de recherche. Si le résultat est négatif, nous utilisons alors le WORDNET pour chercher le synonyme ou le champ lexical du mot clé. La fin de cette phase est couronnée par le vecteur descriptif obtenu grâce à la fonction de recherche.

La dernière phase permet d'obtenir l'index final du document qui commence par une phase de prétraitement qui est l'une des phases les plus importantes dans de nombreuses applications comme la fouille de texte(textmining), le traitement du langage naturel et l'indexation de texte. Essentiellement, comprenant la tokenisation, la suppression des mots vides et la ponctuation, etc. Selon l'organigramme 11, le processus de création de racine inclut à la fois une phase de création basée sur des règles avec une version modifiée de l'algorithme CAS (ajout d'une étape de prétraitement) et une phase d'enrichissement sémantique basé sur l'Ontologie Wordnet. Les résultats obtenus sont comparés et enfin, l'algorithme d'extraction SECAS est appliqué pour obtenir les termes d'index final.

Notre organigramme peut être résumé en trois phases :

*Phase1 : « Acquisition du document »*qui est à son tour, composée de trois étapes :

Étape1 : Le document est donné en entrée

Étape2 : Saisie des principaux descripteurs du document (dit annotation qui est dans notre cas manuelle et incombe au créateur du document). Quatre mots-clés (ou informations clé) sont retenus en plus de l'année et lieu de publication.

Étape3 : sauvegarde des mots clés dans une base de données.

*Phase2 : « Pré-indexation »*qui consiste en :

Étape1 : Application de l'algorithme recherche

Sous étape1 : Recherche dans la folksonomie

Sous etape2 : Identification des descripteurs WordNet

Sous étape3 : Obtention des indexeurs primaux

Phase3 : « Indexation finale » qui consiste en 6 étapes[18] *:*

Étape1 : Suppression de la ponctuation, des chiffres et des mots vides comme préposition, conjonction, article, etc.

Étape 2 : Identification des descripteurs de CAS.

Étape 3 : Identification des descripteurs Wordnet.

Étape4 : Application de l'algorithme SECAS et comparaison des descripteurs

Étape 5 : Identification des descripteurs les plus pertinents.

Étape 6 : Création du fichier d'index final en tant que sortie du stemmer

Pour plus d'illustration nous vous proposons ci-dessous ; le schéma directeur des procédures traitant la phase d'indexation.

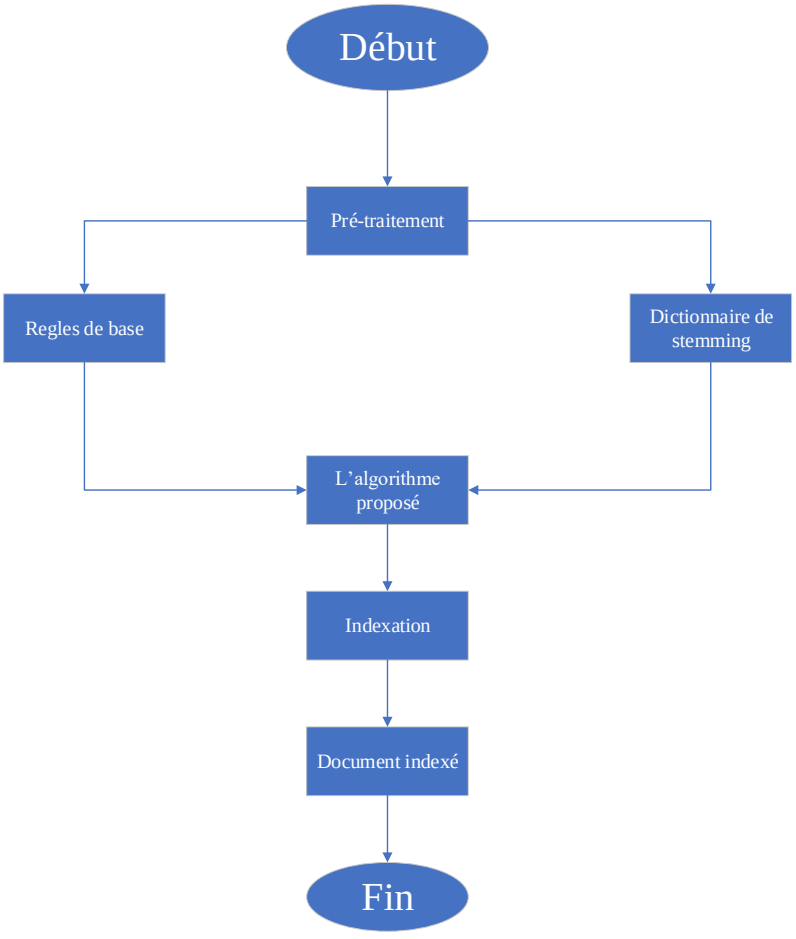

Figure 11: Organigramme de l'indexation

[18] Pour plus d'informations sur cette étape, revenir à la référence [35].

2.2. Les mesures de la similarité

Afin de tester notre méthode de pré-indexation, nous avons utilisé la technique de mapping document-requête présentée dans [35] et qui se base sur une similarité sémantique dite globale car elle-même composée d'une similarité terminologique (à son tour composée d'une similarité lexicale et une similarité syntaxique) et une similarité structurelle se basant sur la structure de la base lexicale Wordnet.

Dans ce qui suit, nous définissons ces mesures utilisées.

- *Similarité terminologique*

Les paires de mots sont comparées et la similarité terminologique est calculée sur la base d'une comparaison syntaxique et lexicale. Les méthodes syntaxiques sont basées sur la comparaison de mots, de chaînes de caractères ou de textes en fonction des lettres qu'ils ont en commun. Les méthodes linguistiques ou lexicales utilisent souvent des ressources externes (dictionnaires, taxonomie, etc.) pour effectuer la comparaison et calculer la similitude entre deux entités, représentées par des termes. Cette similarité est calculée en utilisant les liens sémantiques qui existent dans ces ressources. Dans notre algorithme de mappage, la distance de Jaro [**70**] est utilisée pour la syntaxe similarité et WordNet pour la similarité lexicale.

Par la suite ces deux mesures vont être combinées selon l'équation :

$$Sim_{term}(c_1, c_2) = \frac{(SimLex(c_1, c_2) * CoeffLex) + (Simsyn(c_1, c_2) * CoeffSyn))}{CoeffLex + CoeffSyn}$$

- *Similarité syntaxique*

La distance de Jaro [**70**] prend en compte dans la comparaison de deux chaînes de caractères, d'une part, le nombre de caractères en commun, et aussi l'ordre des lettres (dit transpositions). Il a été prouvé que la distance de Jaro distance fournie une performance intéressante et est plus rapide par rapport à d'autres méthodes de calcul de similarité syntaxique.

Cette mesure est particulièrement adaptée à la comparaison des chaînes courtes et donc parfaite pour mapper les descripteurs de document et les mots-clés de requête lorsque ces dernières sont soumises à une phase d'indexation. Le résultat est

normalisé, de manière à avoir une mesure entre 0 et 1, zéro étant l'absence de similitude.

- *Similarité lexicale*

Les méthodes basées sur un langage se fondent sur des techniques de traitement du langage naturel afin de trouver des associations entre les entités ou les classes. Ces méthodes exigent l'utilisation de ressources externes. Plusieurs types de ressources peuvent être employés, notre choix s'est porté sur WordNet, qui regroupe des termes (noms, verbes, adjectifs et adverbes) en ensembles de synonymes appelés synsets. Un synset regroupe tous les termes dénotant un concept donné.

$$SimLex(c_1, c_2) = \lambda(S)/\min(Syn(c_1), Syn(c_2))$$

Tel que : λ la cardinalité de S est :

$$\lambda(S) = |Syn(c_1) \cap Syn(c_2)|$$

3. Les algorithmes de la solution proposée

Ci-dessous, nous présentons les algorithmes les plus importants correspondants aux différentes tâches de notre algorithme de pré-indexation basé sur les folksonomies.

3.2. Algorithme de vérification des tags

Cet algorithme permet de vérifier si les mots clés choisi lors du chargement du document correspondent à des tags de notre folksonomie. Ces tags seront dans ce cas utilisés comme des indexeurs primaux pour le document afin de permettre sa recherche avant que l'indexation finale ne se fasse. Dans le cas contraire, nous allons essayer de chercher si le mot-clé donné possède un sens (synonyme) dans la ressource wordnet pour voir si ce dernier appartient à la folksonomie.

Algorithme de vérification existence tags
entrée : mc: le mot_clé v[n]: vecteur de tag sortie: exist: booleen début exist: faux

```
 pour v(i)=1 a length(v) faire
si mc=v[i]
exist=vrai
 sinon
wordnet.definition(mc)
fsi
 fin pour
fin
```

4. Algorithme SECAS

L'algorithme SECAS à ce stade est juste donné à titre d'illustration. Utilisé dans la phase d'Indexation finale, ce dernier fait recours à des techniques de stemming et de lemmatisation pour garantir la génération de descripteur sémantiquement correcte se basant sur le contexte lexicale du document.

Algorithme SECAS

```
SECAS (mot)
Entrée : mot (String)
Sortie : Terme indexé (String) représente le meilleur indexe
Début
WordNetTermeindexé ← WordNet Stemming  (mot);
CAS Terme indexé ← CAS (mot)
Terme indexé : vide
Si  mot est un nom composé
Termé indexé ← mot
FinSi
Sinon WordNet terme indexé ≠ mot ou CAS terme indexé ≠ mot
Alors
Si (length (WordNet terme indexé)  <length( CAS terme indexé) ET length (WordNet
terme indexé) > 0)
Alors
Terme indexé ← WordNet terme indexé
Fin
Sinon
CAS terme indexé a au moins une définition
 Alors
CAS terme indexé est plus court que WordNet terme indexé
 Terme indexé ← CAS terme indexé
Fin
Fin
Retourne terme indexé
Fin
```

5. Algorithme de la recherche d'un document

Le pseudo code ci-après, illustre la recherche d'un document. Autrement dit, la fonction d'appariement requête-document.

Algorithme de la recherche d'un document
Entrée : requête Sortie : tableau qui contient la similarité et classement de 10 documents premiers Début Pour Nouveau document Pour jusqu'à la taille de tableau du document o++ / *calculer combien de mot clé */ Nouvelle requête Calculer la similarité sémantique entre chaque mot clé de la requête avec chaque mot clé de document Garder le nombre maximum de similarité Additionner les similarités maximum de requête avec chaque document diviser sur le nombre de mot clé de la requête fin pour obtention tableau des correspondances fin

Remarque :

Pour des fins de traitement et de lisibilité, nous avons aussi été amenées à procéder par des prétraitements et des nettoyages des données. Nous y reparlerons dans le chapitre suivant.

6. Conclusion

Le mérite de l'algorithme SECAS est de réduire la taille des fichiers d'index jusqu'à 60% et d'améliorer le rappel et la précision par rapport aux Algorithmes Porter et Cas. Seulement voilà, le temps d'exécution du processus d'indexation d'une collection donnée est assez onéreux ;

Dans ce chapitre, nous avons présenté notre algorithme de pré-indexation qui consiste à vérifier l'existence des mots clés d'un document existant dans la base des données avec les tags. Ceci aura pour effet de permettre à un document donné de pouvoir être retrouvé aussitôt qu'il est chargé sans attendre que le processus d'indexation sémantique soit déclenché.

Dans le chapitre suivant, nous allons présenter les tests plutôt encourageants que nous avons effectué sur le dataset wiki10+[19].

[19]http://blog.zubiaga.org/2009/11/wiki10-a-wikipedia-based-social-tagging-dataset/

Chapitre V :
Tests et validation du système

1. Introduction

Nous allons à présent présenter l'implémentation de notre solution ainsi que les tests effectués. Dans un premier lieu, nous parlerons de l'environnement de développement (langage et outils). Ensuite, nous parlerons de notre collection de tests ainsi que les résultats obtenus.

2. Collection de données

Dans le but de mener une expérimentation sur notre approche, nous avons collecté un ensemble de documents WIKI10 extrait des sites de Bookworking Social et Wikipedia. Cet ensemble est constitué de 20764 URL, chacune avec ses balises sociales présentées sous format XML.

Pour notre travail, nous avons utilisé deux fichiers, dont le premier présente les documents sur lesquelles nous avons appliqué un algorithme de filtrage qui supprime les informations inutiles et conserve celles qui sont utiles.

« Les figures 12 et 13 » illustrent le document XML avant et après l'application de l'algorithme de nettoyage des données.

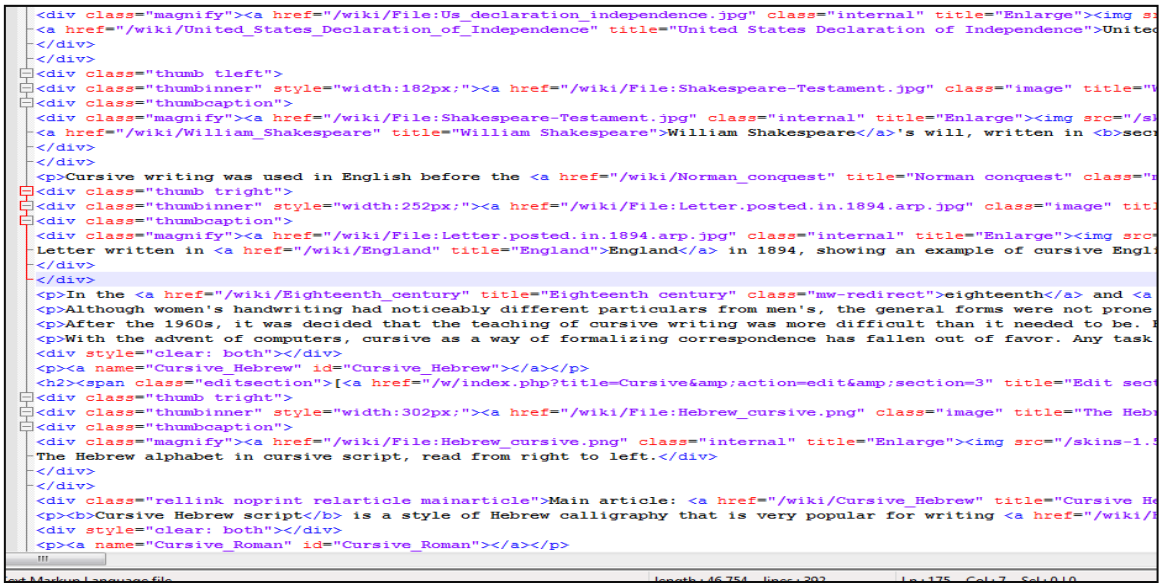

Figure 12: document XML avec balises HTML

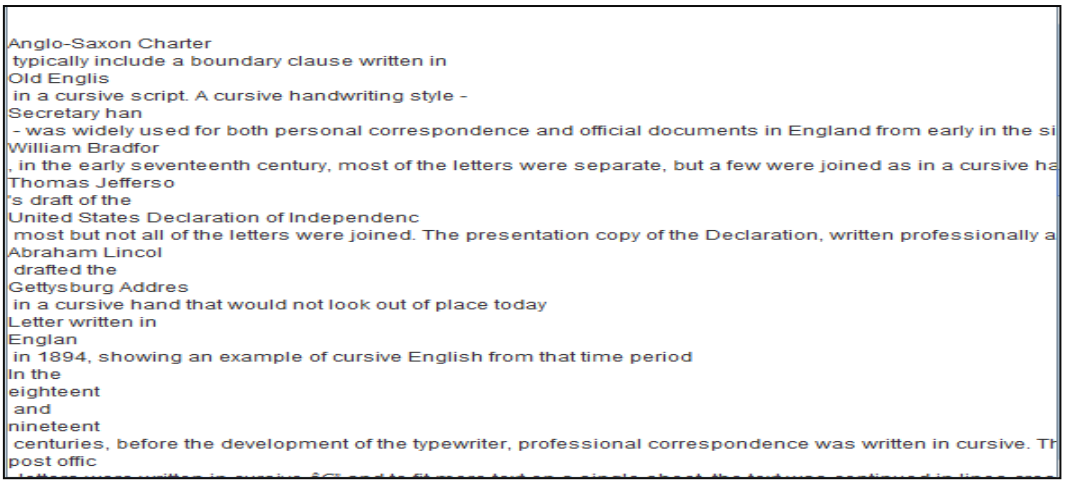

Figure 13: document XML nettoyé.

Le deuxième fichier utilisé, présente l'ensemble des Tags dont chacun est identifié par le nom et le count. Nous avons utilisé un algorithme qui supprime les balises, ce qui nous permet de récupérer seulement les informations essentielles (dans notre cas nom, count).

3. Technique d'évaluation

Les mesures utilisées pour évaluer la qualité des correspondances produites entre les mots clés d'une requête qu'on appellera **R** et le document tagué ou pré-indexé qu'on notera **D** sont principalement les mesures de calcul de la pertinence en recherche d'information, telle que la précision et le rappel.

- ***TP :*** Les correspondances *correctes trouvées* par un système sont appelées« *the true positives (TP)* » et sont calculées ainsi :

$$TP=D\cap R$$

- ***FP :*** Les correspondances *incorrectes trouvées* par un système sont appelées « the false positives (FP) » et sont calculées ainsi :

$$FP=D- D\cap R$$

- ***FN :*** Les correspondances *correctes omises* par un système sont appelées« the false négatives (FN) » et sont calculées ainsi :

$$FN=R- D\cap R$$

- La précision est une mesure d'exactitude, elle varie entre [0,1], elle est calculée de la manière suivante.

$$\text{Précision} = \frac{|TP|}{|TP+FP|}$$

- Le rappel est une mesure de perfection, elle varie entre [0,1], elle est calculée de la manière suivante :

$$\text{Rappel} = \frac{|TP|}{|TP+FN|}$$

4. Environnement de développement

4.1. Langages utilisés

5.1.4. Java

Java est un langage de programmation informatique orienté objet créé par James Gosling et Patrick Naughton de Sun Microsystems. Il permet de créer des logiciels compatibles avec de nombreux systèmes d'exploitation (Windows, Linux, Macintosh, Solaris). Le langage Java donne aussi la possibilité de développer des programmes pour téléphones portables et assistants personnels. Enfin, ce langage peut-être utilisé sur internet pour des petites applications intégrées aux pages web (applet) ou encore comme langage serveur (JSP) [**71**].

5.1.5. MYSQL

MYSQL est un Système de gestion de bases de données relationnelles (SGBDR)sous licence GNU très utilisé pour mettre en ligne des bases de données.Il permet d'entreposer des données de manière structurée (Base, Table, Champs,Enregistrements). Le noyau de ce système permet d'accéder à l'informationentreposée via un langage spécifique le SQL**.**

4.2. OUTILS

5.1.6. Eclipse

Eclipse IDEest un environnement de développement intégré libre, le terme *Eclipse* désigne également le projet correspondant, lancé par IBM. Il est extensible, universel et polyvalent, permettant potentiellement de créer des

projets de développement mettant en œuvre n'importe quel langage de programmation. Eclipse IDE est principalement écrit en Java à l'aide de la bibliothèque graphique SWT, d'IBM [**72**] [**73**].

4.3. Les APIS

- ***Swing :*** est une bibliothèque graphique pour le langage de programmation Java, faisant partie du package Java Foundation Classes (JFC), inclus dans J2SE. Swing constitue l'une des principales évolutions apportées par Java 2 par rapport aux versions antérieures. Et il offre la possibilité de créer des interfaces graphiques identiques quel que soit le système d'exploitation sous-jacent, au prix de performances moindres qu'en utilisant Abstract WindowToolkit (AWT). Il utilise le principe Modèle-Vue-Contrôleur (MVC, les composants Swing jouent en fait le rôle de la vue au sens du MVC) et dispose de plusieurs choix d'apparence pour chacun des composants standards.
- ***HTML parser :***Ce module définit une classe HTMLParser qui sert de base à l'analyse de fichiers texte au format HTML (HyperText Mark-upLanguage) et XHTML.

5. Présentation des test

5.1. Processus de nettoyage et de pré-indexation

La décision de recherche est prise en comparant les termes de la requête avec les termes de Tags (mots ou expressions importants) apparaissant dans le document lui-même; la décision peut être binaire (récupérer/rejeter), ou impliquer l'estimation de la pertinence, degré de similarité entre le document et la requête. Ainsi, avant que les documents ne soient insérés, des techniques de prétraitement des données sont appliquées à la collection afin de réduire sa taille en supprimant autant que possible les variantes structurelles de mots ayant est mêmes significations. Cette action augmenterait l'efficacité du système RI.

- ***Extraction des mots clés :*** « La tokenisation est le processus qui consiste à séparer les phrases ainsi que le fichier texte en mots délimités par un espace, une tabulation ou une nouvelle ligne ». En d'autres termes, la tokénisation est le processus de conversion d'un flux de caractères (le texte des documents) en un flux de mots (les mots candidats à

adopter en tant que termes d'index), c'est-à-dire l'identification des différents mots du texte.

- ***La suppression des mots vides :***Un mot vide est un mot qui n'a aucune signification en soi. Les mots les plus courants représentent environ 50% du contenu du texte dans la plupart des langues. Ces mots ne sont pas utiles pour décrire le besoin d'information de l'utilisateur. Il en va de même pour les déterminants, les conjonctions de coordination, les prépositions, les articles, les verbes auxiliaires, les pronoms relatifs, etc. Ainsi, utiliser ces mots ne ferait que ralentir l'IRS sans améliorer, ni la phase d'indexation ni la pertinence des résultats retournés.
- ***Suppression de la ponctuation :***Afin d'obtenir de bons résultats et d'accélérer le processus d'empilement, il est également important de supprimer toutes les ponctuations ainsi que les accents inclus dans le contenu des textes, car ils sont dépourvus de sens et non pertinents pour une tâche de recherche donnée.

5.2. Les Interfaces de l'application

- *Interface d'ajout :* On a fait un formulaire pour que l'utilisateur puisse ajouter un document (figure 14). Ceci est l'interface principale de l'application:

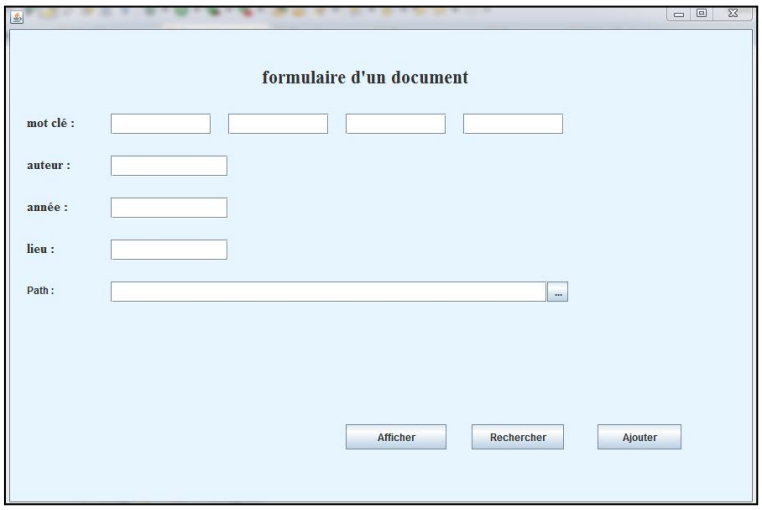

Figure 14: Le formulaire d'ajout d'un document.

Cette dernière permet à l'utilisateur de saisir les informations du document après l'avoir choisir grâce au bouton file chooser qui permet de choisir un fichier dans le disque dur de l'ordinateur.

- *L'interfacce d'Affichage :* Elle permet d'afficher un document XML sans balises « HTML » et le nettoyer pour n'avoir seulement le texte qui s'affiche.

Un clic sur « Affichage » fait apparaître la fenêtre suivante :

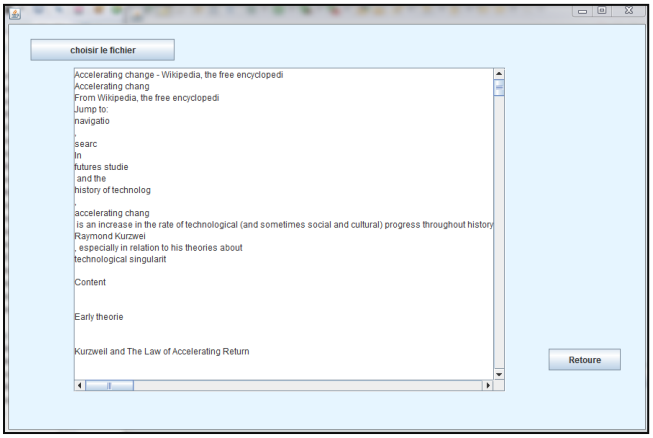

Figure 15: affichage d'un document sans balises HTML

- *L'interface de recherche :* Un clic sur *« Rechercher »* permet d'afficher la fenêtre présentée dans la Figure 16. Cette dernière permet à saisir une requête quelconque. Ensuite, les mots clés de la requête sont extraites comparés avec les mots clés tags des documents, présents dans le pré-indexe afin de calculer la similarité avec les différents documents du corpus.

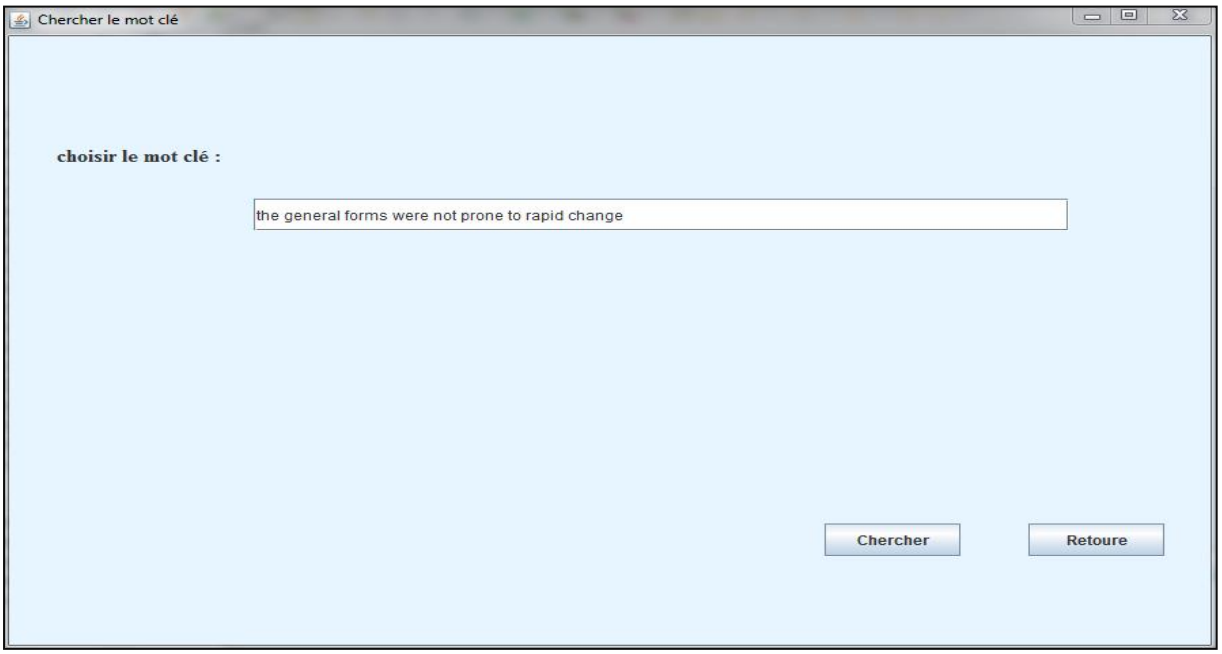

Figure 16: saisir une requête et extraire ses mots clés

- *Cette fonction* permet de chercher les mots clés entre requête saisie et les documents dans la table des documents tagués et à la fin on calcule la similarité de chaque document par rapport à la requête et le classé.

```
taille avant traitement:2
taille aprÃ¨s traitement:2
Loading modules
set up:
```

Figure 17: résultat d'un extraire des mots clés d'une requête

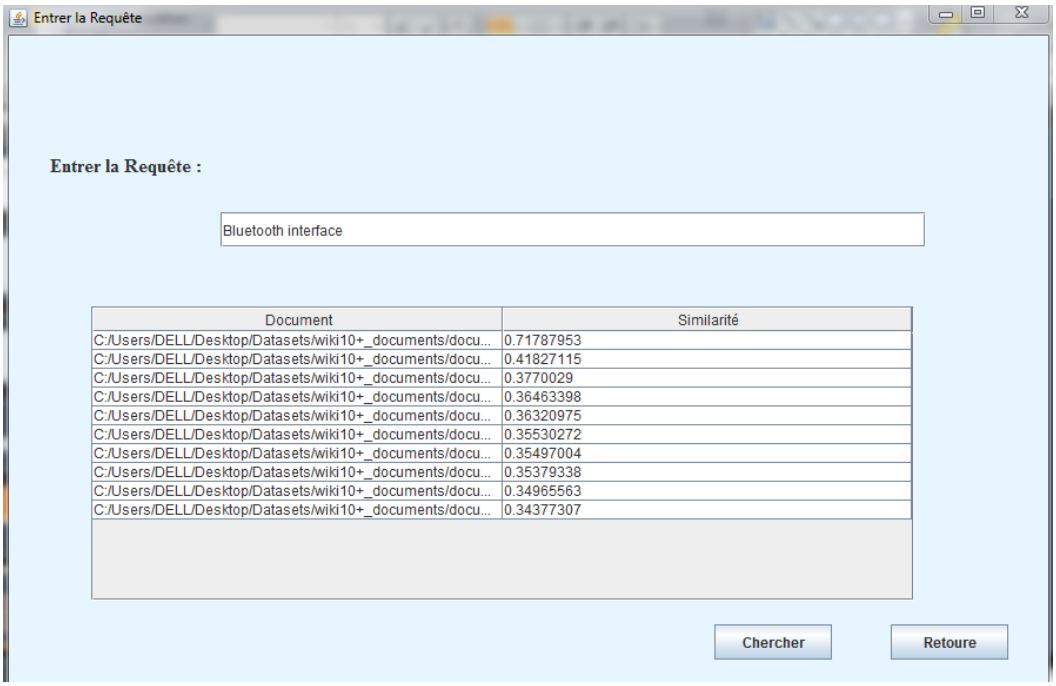

Figure 18: affichage d'un document et son similarité

5.1. Base de donnée

Pour nos tests, nous avons mis en place une petite base de données, représentée dans les deux tableaux suivants. Ses tables contiennent toutes les informations relatives à la Folksonomies, les documents, leurs tags (pré-indexe) et éventuellement la version indexée dans le future.

id-docum ▲ 1	mcc1	mcc2	mcc3	mcc4	Auteur	Annee	Pays	path-doc
1	technical	API writer	aerospace	association for business communication	wikipedia	2009	united state	000e9edf0163688ef62a4592546109fb.html
2	lian banks	special search	Anarchism	Admiralty	wikipedia	2009	united state	00a9f3c4147462e18229d68479ad7784
3	abuer	Albert Einstein	CHSH	Complementarity	Wikipedia	2009	united state	00a047fae28b588e3caae1dea08179a0
4	accelerating	Aging research	Agricultural revolution	Anthropology	Wikipedia	2009	united state	00a634f5bbba01e2ccd1e2e3d001dd9c
5	cursive	Anglo-Saxon Charters	Alphabet	Arabic	Wikipedia	2009	united state	00a1446e1cdea374f8324a6ee72cbb1b
6	Land art	Alan Sonfist	Algae	Arizona	wikipedia	2009	united state	00b2bd53f3daf60170847506acec6488
7	Socialist	Articles unsourced statements	Abramtsevo Colony	Aleksandr Deineka	wikipedia	2009	united state	00b5bb50edd9283e3f6702039e15e90b
8	Honey and Clover	United States	Germany	Indonesia	wikipedia	2009	united state	00b94b1c063dc764a7ae70d44ab8cd36
9	IPAQ	ARM architecture	Bluetooth	Cannes	wikipedia	2009	united state	00b170e56b73914304d4e97bf62e5613
10	Search in literature	Ada Dietz	Aeneid	Alexander Girard	wikipedia	2009	united state	00b827ae362bdbe86190035bee58e1b0.htm
11	Animal testing	Asphyxia	BBC News Online	Baikonur Cosmodrome	wikipedia	2009	united state	00bc3b3d3d5da747355b797db01c242e
12	Advantage	Barona Casino	Blackjack Forum	Card counting	wikipedia	2009	united state	00be38c68f94c17d0eedc24dcd46e818
13	Compression Software Implementations	Russian language text	Compression formats	ALLPlayer	wikipedia	2009	united state	00c4cc1519028443658a3f0a0611f5e4
14	Articles needing additional references	Fact	12th Street riot	Democratic National Convention	wikipedia	2009	united state	00c446f54dbb6c1557845eef7b93940b
15	Human brain	Footer Neuropsychology	Arousal	Attention	wikipedia	2009	united state	00c683c0b6f294ae69abb717e98f2f0f

Tableau 4: Table document dans la base de données.

Ce dernier tableau représente la table des documents ainsi que leurs mots clés, auteur, pays, année et le path qui permet d'accéder au document (tableau 4).

L'autre tableau représente les documents pré-indexéspar les tags de la folksonomie (tableau 5)

id-docpre	id-docum ▲ 1	path-doc	nbr-occ
41	1	000e9edf0163688ef62a4592546109fb.html	4
42	2	00a9f3c4147462e18229d68479ad7784	4
43	3	00a047fae28b588e3caae1dea08179a0	3
44	4	00a634f5bbba01e2ccd1e2e3d001dd9c	4
45	5	00a1446e1cdea374f8324a6ee72cbb1b	4
46	6	00b2bd53f3daf60170847506acec6488	4
47	7	00b5bb50edd9283e3f6702039e15e90b	4
48	8	00b94b1c063dc764a7ae70d44ab8cd36	4
49	9	00b170e56b73914304d4e97bf62e5613	4
50	10	00b827ae362bdbe86190035bee58e1b0.html	4
51	11	00bc3b3d3d5da747355b797db01c242e	3
52	12	00be38c68f94c17d0eedc24dcd46e818	4
53	13	00c4cc1519028443658a3f0a0611f5e4	3
54	14	00c446f54dbb6c1557845eef7b93940b	4
55	15	00c683c0b6f294ae69abb717e98f2f0f	4
56	16	00c741fb4cf489c30a74c9b72917b87d	4
57	17	00ce4dcc93e6d1f4de32ab943c14b696	4
58	18	00ce09e91b355a931140b503c6537eeb	4
59	19	00d58c9caf949240e0d433256d330c44	3
60	20	00d737c4c01e4d1b42212c83518a4938	4

Tableau 5: Table des documents pré-indexée

Cette table contient les documents ainsi que les mots-clé retenus après pré-indexation et comparaison avec la Folksonomie.

5.1.7. Diagramme de classe

Le schéma ci-après représente une illustration des tables (classes du système). Où la classe « Documents » contient tous les documents du corpus, la classe « DocPreindexe » contient les informations de pré-indexation et la classe « DocIndexé » représente l'index final de la collection.

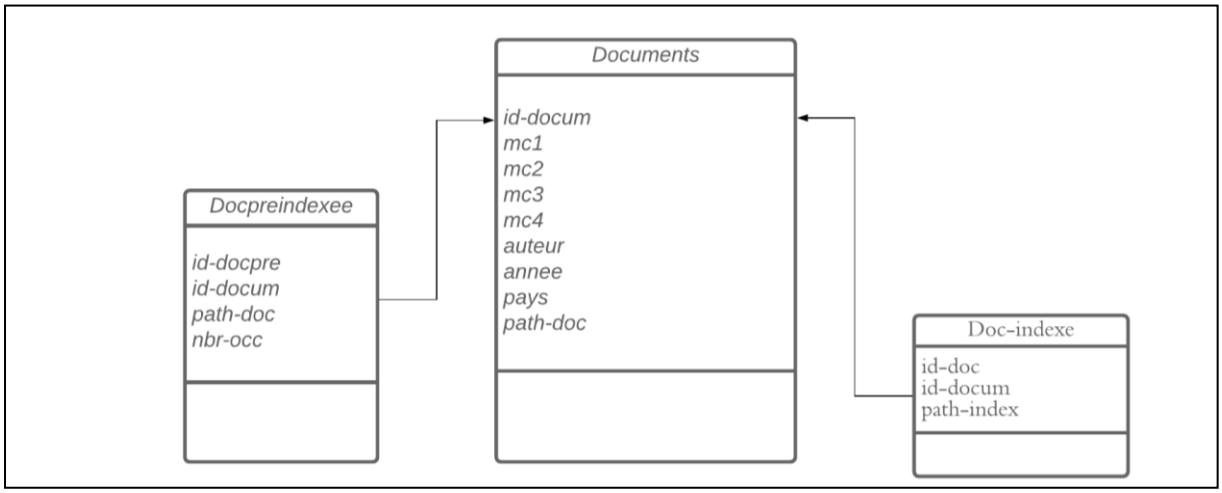

Figure 19: notre base de données

6. Les tests et validation

6.1. Les requêtes proposées

Afin de tester nos algorithmes, nous avons élaboré des requêtes variées présentée dans le tableau ci-dessous.

Nom	Requêtes
Req1	Technical communication in germany
Req2	Iain Banks and science fiction
Req3	Qur'an in Classical Arabic

Req4	art in Indonesia
Req5	the first work of socialist realism in land art
Req6	what is the difference between indexation and folksonomy
Req7	Bluetooth interface
Req8	CHSH inequality and Complementarity
Req9	The Law of Accelerating and Aging research
Req10	art and literature

Tableau 6:les requetes proposées

6.2. Table de correspondances

Le tableau suivant montre un échantillon des correspondances manuelles et correspondances basée folksonomie entre les documents tagués (10 documents retenus) dans la base de données et une requête donnée avec un seuil = 0,50 (le tableau entier sera présenté en annexes).

Remarques :

Après quelques tests, nous avons opté pour le seuil 0.5 car il donne les meilleurs résultats d'appariement à notre goût.

C'est vrai que si on compare aux méthodes d'indexation classiques, nous trouverons que le seuil est relativement petit, mais on peut expliquer ça du fait que dans notre indexation, nous n'avons pas pris en considération le contenu textuel dans sa globalité, mais nous nous sommes uniquement intéressé aux mots clés saisie lors du chargement du fichier (qui peut être assimilé à l'action de tag dans les réseaux sociaux). Ceci nous a obligé de choisir un seuil stricte qui a toutefois montré de bons résultats tels que les bon documents étaient toujours relativement bien classé.

Requête	Document	Correspondances manuelles	Mesures de similarité	Correspondances De folksonomie
Req5	Document7	√	0.542	√
Req5	Document20	×	0.53615	√
Req5	Document15	×	0.53610	√
Req5	Document6	√	0.618	√
Req5	Document2	√	0.526	√
Req5	Document10	×	0.525	√
Req5	Document14	×	0.517	√
Req5	Document19	×	0.487	×
Req5	Document8	√	0.48470	×
Req5	Document16	×	0.48410	×

Tableau 7: tableau test documents / requête

6.3. Mesures de performance

A travers le tableau qui suit, nous pouvons constater que mesures de performances de Notre système sont plutôt encourageantes.

TP	FP	Précision	Rappel	*F-mesure*	Overall	Fallout
20	16	0.55	0.58	0.56	0.11	0.44

Tableau 8: Mesures de performances du folksonomie

6.4. Analyse des résultats

Nous remarquons que le rappel et la précision sont légèrement en dessus de la moyenne ce qui est, même si ce n'est pas énorme, plutôt encourageant. De plus, La F-mesure qui donne une vision globale de performance du système est plutôt harmonieuse et conforme aux « rappel » et « précision ».

Le overall n'a pas pris de valeurs négatives ce qui témoigne que le système ne donne pas de FP ce qui est rassurant, il n'est tout fois pas supérieur à 0.5 et ceci est sûrement du au non traitement du contenu textuel qui aurait pu améliorer ses valeurs.

Pour finir le pourcentage d'erreur est assez faible aussi.

7. Conclusion

Dans ce chapitre nous avons présenté l'implémentation de notre solution ainsi que la collection de tests que nous avons utilisés et les résultats obtenus. A ce titre, nous avons observé que les résultats obtenus étaient plutôt encourageants et mériteraient d'être poursuivis et enrichis.

Conclusion générale

8 . Conclusion générale

La Recherche d'Information (RI) est incontournable dans les sociétés modernes et elle est basée sur le savoir. Les environnements d'information sont de plus en plus nombreux, complexes et omniprésents, car la quantité d'informations hétérogènes disponibles augmente de façon exponentielle chaque année.

La recherche d'information sur les données (dataset) est gourmande en temps et d'espace mémoire, En outre la recherche d'information est souvent imprécise parce que la taille d'information est énorme, il est bien connu que l'indexation est la tâche la plus onéreuse de la RI. Le but principal de l'indexation est de créer une représentation des documents présents dans le corpus de documents global de façon à automatiser les traitements de ces derniers et faciliter leur appariement avec les requêtes des utilisateurs.

Dans ce contexte, la recherche sémantique a été introduite dans le but d'améliorer la recherche et d'éclaircir le sens contextuel des termes. En plus du problème sémantique, RI rencontre de nombreux problèmes qui ont été présenté dans notre mémoire.

Dans ce projet, nous nous sommes intéressées à l'étude de l'utilisation des Floksonomies comme moyen rapide et personnalisé de représenter et d'indexer l'information car considérées comme l'alternative moderne aux Ontologies et ressources lexicales traditionnellement utilisées dans la RI. D'autre part, le but est d'analyser si cette pratique peut améliorer l'efficacité d'une méthode classique.

Dans un premier temps, notre travail consiste enla vérification des documents qui sont tagués.Cette opération est réalisée à l'aide d'un algorithme qui contient les fonctions de nettoyage des documents pour récupérer que les informations nécessaires et réduire la taille de document. Elle permet aussi de réduire le temps d'obtention des résultats recherchés. Nous avons également associé le wordNet pour la recherche des synonymes des mots qui ne figurent pas dans les tags. Pour finir, nous avons aussi calculé la similarité entre les requêtes introduites et les documents tagués issus du corpus Wiki10+.

Nous avons à cet effet remarqué que les résultats étaient plutôt encourageants, mais qu'il fallait toute fois reconduire d'autres test avec d'autres collections peut être et surtout comme perspectives de ce travail, il faut essayer de voir du côté des social

media où le concept de tags prends son sens et où l'intéré de trouver les informations en temps réel est critique.

Annexes

Annexe A : Table des correspondances requêtes-documents.

Le tableau ci-dessous, représente toutes les correspondances des tests que nous avons effectués avec les 10 requêtes formulés dans le chapitre 4 avec 10 des documents de la collection de tests utilisée.

Requêtes	**Documents**	**Correspondances manuelles**	**Mesures de similarité**	**Correspondances De folksonomie**
Req1	Doc1	√	0.67	√
	Doc2	×	0.46	×
	Doc3	√	0.37	×
	Doc4	×	0.34	×
	Doc5	×	0.37	×
	Doc6	×	0.56	√
	Doc7	×	0.38	×
	Doc8	√	0.58	√
	Doc9	×	0.40	×
	Doc10	×	0.47	×
	Doc11	×	0.23	×
	Doc12	×	0.34	×
	Doc13	×	0.26	×
	Doc14	×	0.40	×

	Doc15	×	0.50	√
	Doc16	×	0.43	×
	Doc17	×	0.35	×
	Doc18	×	0.41	×
	Doc19	×	0.45	×
	Doc20	×	0.45	×
Req2	Doc1	×	0.39	×
	Doc2	×	0.39	×
	Doc3	×	0.39	×
	Doc4	×	0.41	×
	Doc5	×	0.40	×
	Doc6	×	0.45	×
	Doc7	×	0.38	×
	Doc8	×	0.36	×
	Doc9	×	0.30	×
	Doc10	×	0.46	×
	Doc11	×	0.24	×

	Doc12	×	0.35	×
	Doc13	×	0.26	×
	Doc14	×	0.46	×
	Doc15	√	0.46	×
	Doc16	×	0.36	×
	Doc17	√	0.53	√
	Doc18	×	0.38	×
	Doc19	×	0.39	×
	Doc20	×	0.44	×
Req3	Doc1	×	0.33	×
	Doc2	×	0.36	×
	Doc3	×	0.31	×
	Doc4	×	0.29	×
	Doc5	√	0.58	√
	Doc6	√	0.42	×
	Doc7	×	0.33	×
	Doc8	×	0.28	×

Annexe A : Table des correspondances requêtes-documents.

	Doc9	×	0.29	×
	Doc10	×	0.38	×
	Doc11	×	0.25	×
	Doc12	×	0.31	×
	Doc13	×	0.24	×
	Doc14	×	0.39	×
	Doc15	×	0.32	×
	Doc16	×	0.32	×
	Doc17	×	0.34	×
	Doc18	√	0.37	×
	Doc19	×	0.28	×
	Doc20	√	0.34	×
Req4	Doc1	√	0.52	√
	Doc2	×	0.44	×
	Doc3	×	0.42	×
	Doc4	×	0.38	×
	Doc5	×	0.39	×

	Doc6	√	0.81	√
	Doc7	×	0.33	×
	Doc8	√	0.68	√
	Doc9	×	0.51	√
	Doc10	×	0.47	×
	Doc11	×	0.25	×
	Doc12	×	0.38	×
	Doc13	×	0.27	×
	Doc14	×	0.44	×
	Doc15	×	0.49	×
	Doc16	×	0.44	×
	Doc17	×	0.33	×
	Doc18	×	0.38	×
	Doc19	×	0.58	√
	Doc20	×	0.45	×
Req5	Doc1	×	0.61	√
	Doc2	√	0.52	√

Annexe A : Table des correspondances requêtes-documents.

	Doc3	×	0.46	×
	Doc4	×	0.41	×
	Doc5	×	0.40	×
	Doc6	√	0.61	√
	Doc7	√	0.54	√
	Doc8	√	0.48	×
	Doc9	×	0.43	×
	Doc10	×	0.52	√
	Doc11	×	0.26	×
	Doc12	×	0.41	×
	Doc13	×	0.27	×
	Doc14	×	0.51	√
	Doc15	×	0.53	√
	Doc16	×	0.48	×
	Doc17	×	0.45	×
	Doc18	×	0.47	×
	Doc19	×	0.48	×

	Doc20	×	0.53	√
Req6	Doc1	×	0.46	×
	Doc2	√	0.61	√
	Doc3	×	0.35	×
	Doc4	×	0.40	×
	Doc5	×	0.43	×
	Doc6	√	0.54	√
	Doc7	×	0.34	×
	Doc8	×	0.39	×
	Doc9	×	0.31	×
	Doc10	×	0.53	√
	Doc11	×	0.26	×
	Doc12	×	0.37	×
	Doc13	×	0.24	×
	Doc14	√	0.45	×
	Doc15	×	0.54	√
	Doc16	√	0.62	√

	Doc17	×	0.41	×
	Doc18	×	0.42	×
	Doc19	×	0.36	×
	Doc20	×	0.48	×
Req7	Doc1	×	0.41	×
	Doc2	×	0.30	×
	Doc3	√	0.34	×
	Doc4	×	0.35	×
	Doc5	×	0.36	×
	Doc6	×	0.35	×
	Doc7	×	0.30	×
	Doc8	×	0.32	×
	Doc9	√	0.71	√
	Doc10	×	0.37	×
	Doc11	×	0.23	×
	Doc12	×	0.29	×
	Doc13	×	0.25	×

	Doc14	×	0.36	×
	Doc15	×	0.34	×
	Doc16	×	0.35	×
	Doc17	×	0.33	×
	Doc18	√	0.34	×
	Doc19	×	0.32	×
	Doc20	×	0.31	×
Req8	Doc1	×	0.35	×
	Doc2	×	0.41	×
	Doc3	√	0.34	×
	Doc4	×	0.33	×
	Doc5	√	0.59	√
	Doc6	×	0.41	×
	Doc7	×	0.35	×
	Doc8	×	0.34	×
	Doc9	×	0.31	×
	Doc10	×	0.43	×

	Doc11	×	0.29	×
	Doc12	√	0.44	×
	Doc13	×	0.26	×
	Doc14	×	0.40	×
	Doc15	×	0.40	×
	Doc16	×	0.38	×
	Doc17	√	0.44	×
	Doc18	×	0.50	√
	Doc19	×	0.32	×
	Doc20	×	0.41	×
Req9	Doc1	×	0.41	×
	Doc2	×	0.43	×
	Doc3	×	0.28	×
	Doc4	√	0.40	×
	Doc5	×	0.44	×
	Doc6	√	0.44	×
	Doc7	×	0.29	×

	Doc8	×	0.32	×
	Doc9	×	0.28	×
	Doc10	×	0.50	√
	Doc11	×	0.22	×
	Doc12	×	0.30	×
	Doc13	×	0.25	×
	Doc14	√	0.47	×
	Doc15	√	0.47	×
	Doc16	×	0.45	×
	Doc17	×	0.39	×
	Doc18	×	0.34	×
	Doc19	×	0.32	×
	Doc20	×	0.38	×
Req10	Doc1	×	0.54	√
	Doc2	√	0.61	√
	Doc3	×	0.44	×
	Doc4	×	0.54	√

	Doc5	×	0.51	√
	Doc6	√	0.79	√
	Doc7	×	0.31	×
	Doc8	×	0.38	×
	Doc9	×	0.35	×
	Doc10	√	0.79	√
	Doc11	×	0.23	×
	Doc12	×	0.39	×
	Doc13	×	0.26	×
	Doc14	√	0.53	√
	Doc15	×	0.59	√
	Doc16	×	0.64	√
	Doc17	×	0.41	×
	Doc18	×	0.44	×
	Doc19	×	0.43	×
	Doc20	×	0.48	×

Tableau 9: Table des correspondances requêtes-documents

Références bibliographiques

Références bibliographiques

[1] Boubekeur-Amirouche-Fatiha, "Contribution à la définition de modèles de recherche d'information flexibles basés," TOULOUSE, thèse du doctorat 2008.

[2] Boubekeur F, ""“Contribution à la définition de modèles de recherche d'information", thèse de doctorat en informatique," Université Paul Sabatier, 2008.

[3] Bookstein, "bookstein A. outligne of a general probabilistic reterieval model, Journal of documentation," 1983.

[4] B Ribeiro-Neto and R Baeza-Yates, "Modern information retrieval: The concepts and technology behind search," 2011.

[5] A.W. Woods, ""“Conceptual indexing: A better way to organize knowledge”. Technical Report SMLI TR-97-61, Sun Microsystems Laboratories, Mountain View, CA," ww.sun.com/research/techrep/1997/abstract-61.html, April 1997.

[6] Developpez.com. [Online]. https://jplu.developpez.com/tutoriels/web-semantique/introduction/

[7] Marron et Al, "Marron, M. and Kuhns J. on relevance ,probabilistic indexing and information reterieval. journal of the ACM," 1960.

[8] Robertson Al., "s.robertson, m.marron, and w.cooper probability of revelance: a unification of two competing models for document reterieval. information technology: research and development," 1982.

[9] M. Hammache Arezki, "Recherche d’Information : un modèle de langue combinant mots simples et mots composés," Tizi-Ouzou, Thèse de Doctorat en Informatique.

[10] Ismail BADACHE, "Recherche d’Information Sociale "Exploitation des Signaux Sociaux pour Améliorer la Recherche d'information"," TOULOUSE, thèse de doctorat Février 2016.

[11] G. Salton and M. McGill, ""“Introduction to Modern Information Retrieval”," McGraw-Hill, New York, 1983.

[12] Anita Largouet, ""“La recherche d'informations sur Internet. Rapport de recherche”," Service Commun de Documentation - Université Michel de Montaigne - Bordeaux3, 2005.

[13] Jérôme Bondu, ""“Panorama d’outils de recherche d’informations gratuits et en ligne”.," Inter-Ligere Sarl, http://www.inter-ligere.com/article-30587376.html.

[14] Thomas Mandl, "Recent developments in the evaluation of information retrieval systems: Moving towards diversity and practical relevance," Informatica, 32(1) 2008.

[15] Wassila Azzoug, "Contribution à la définition d'une approche d'indexation sémantique de documents textuels," université M'hamed Bougara, Boumerdes-algerie, Mémoire de magister 2012/1013.

Références bibliographiques

[16] J. Hendler et O. Lassila. T. Berners-Lee, ""The semantic web". Scientific American," amérique , 2001.

[17] M.Baziz, ""Indexation conceptuelle guidée par ontologie pour la recherche," Université Paul Sabatier de toulouse, 2005, 05-a.

[18] F. Verdejo, I. Chugur , J. Cigarrán J. Gonzalo, ""Indexing with WordNet synsets can improve text retrieval». in Proc. the COLING/ACL '98 Workshop on Usage of WordNet for Natural Language Processing," 1998.

[19] M. Sanderson, ""Word Sense Disambiguation and Information Retrieval". Technical Report (TR-1997-7) of the Department of Computing Science," University of Glasgow, Glasgow G12 8QQ, UK, 1997.

[20] W. B. Croft. R. Krovetz, ""Lexical Ambiguity and Information Retrieval". ACM Transactions on Information Systems, Vol. 10, No 2, pp. 115_141," April 1992.

[21] R. Krovetz, ""Homonymy and polysemy in information retrieval". In Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics (A CL-97}, pages 72-79," 1997.

[22] I. Ponte, J.M., & Croft, W.B. Peters, "Folksonomies (2009) : indexing and retrieval in Web 2.0. Walter A language modeling approach to information retrieval. In Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval (pp.275-281).," 1998.

[23] Mlle Isra HAMADA, "Utilisation de "WordNet" pour indexation sémantique & recherche d'information," Mémoire de MASTER 2012/2013.

[24] M. R.(1968).Semantic Memory. In M. Minsky (Ed.), Quillian, "Semantic Information Processing (pp.227-270). MIT Press.,".

[25] D.L. American psychologist Medin, "Concepts and conceptual structure. 44(12)," Americ, 1989.

[26] Carlo Abi Chahine, "Indexation et recherche conceptuelles de documents pédagogiques guidées par la structure de Wikipédia," Octobre 2011.

[27] JohnF. Sowa, "Conceptual graphs for a data base interface. IBM Journal of Research and Development, 20(4),336-357.," 1976.

[28] J.F. Sowa, "Conceptual structures : information processing in mind and machine. Addison Weshley.," 1983.

[29] T.R. Gruber, "A translation approach to portable ontology specifications.Knowledge acquisition, 5(2).doi :10.1006/knac.1008," 1993.

[30] N.,& Giaretta,P. Guarino, "Ontologies and knowledge bases : Towards a terminological clarification. Towards very large knowledge bases : knowledge building and knowledge sharing, 1(9).," 1995.

Références bibliographiques

[31] G.A. Miller, "WordNet : a lexical database for English. Communications of the ACM, 38(11)," 1995.

[32] A. Fogarolli, "WordNet : An electronic lexical database. The MIT press. Word sense disambiguation based on wikipedia link structure.," In proceedings of the 2009 IEEE International Conference on Semantic Computing 2009.

[33] T. Vander Wal, "Folksonomy Definition and Wikipedia.Off the Top [blog].," http://www.vanderwal.net/random/entrysel.php?blog=1750 January 3, 2007.

[34] MELYARA MEZZI, "CONTEXT-AWARE INFORMATION RETRIEVAL SYSTEMS: CONTRIBUTION TO A SEMANTICALLY ENRICHED, FOLKSONOMY-BASED TEXT-SEARCH," DOCTORAL THESIS IN COMPUTER SYSTEMS ENGINEERING 2018.

[35] I. and W. G. Stock Peters, "Folksonomy and Information Retrieval. The American Society for Information Science and Technology. 44," 2008.

[36] K., et al Weller, "Social Interaction Technologies and Collaboration Software. Germany," 2010.

[37] J. Trant, "Studying Social Tagging and Folksonomy: A Review and Framework," Canada, 2007.

[38] P Merholz, "Metadata for the Masses.," Retrieved December 15, 2007 from http://www.adaptivepath.com/publications/essays/archives/000361.php 2004b, October 19, 2004.

[39] Susan Leigh Star, "Slouching towards infrastructure. Social Aspects of Digital Libraries Workshop.," Retrieved December 27 from https://is.gseis.ucla.edu/research/dl/star.html February 16-17, 1996.

[40] E Quintarelli, "Folksonomies: power to the people. ISKO Italy - UniMIB Meeting," Milan, http://www.iskoi.org/doc/folksonomies.htm Retrieved December 13, 2005.

[41] C Shirky, "Folksonomies are a forced move: a response to Liz. Many 2 Many [blog]," Retrieved January 31, 2008 from http://many.corante.com/archives/2005/01/22/folksonomies_are_a_forced_move_a_response_to_liz.php 2005a.

[42] L Rosenfeld, "Folksonomies? How about Metadata Ecologies?," Retrieved December 22, 2005 from http://www.louisrosenfeld.com/home/bloug_archive/000330.html January 6, 2005.

[43] E Kroski, "The Hive Mind: Folksonomies and User-Based Tagging. InfoTangle [blog]," Retrieved December 12 from http://infotangle.blogsome.com/2005/12/07/the-hive-mind-folksonomies-and-userbased- 2005.

[44] E Peterson, "Beneath the Metadata: Some Philosophical Problems with Folksonomy. D-Lib Magazine, 12(11).," Retrieved August 7, 2008 from http://www.dlib.org/dlib/november06/peterson/11peterson.html 2006.

Références bibliographiques

[45] D Weinberger, "Beneath the Metadata - a reply. Joho the Blog [blog]," Retrieved November 20. 2007 from http://www.hyperorg.com/blogger/mtarchive/beneath_the_metadata_a_reply.html 2006a.

[46] D Weinberger, "Why Tagging Matters. Berkman Center for Internet & Society," Retrieved November 20, 2007 from http://cyber.law.harvard.edu/home/uploads/507/07-WhyTaggingMatters.pdf 2005b, May 2005.

[47] D Weinberger, "Everything Is Miscellaneous: The Power of the New Digital Disorder: Times Books," 2007.

[48] C Shirky, "Folksonomy. Many2Many [blog]," Retrieved December 11 2005 from http://www.corante.com/many/archives/2004/08/25/folksonomy.php 2004, August 25, 2004.

[49] P Speroni di Fenizio, "On Tag Clouds, Metric, Tag Sets and Power Laws. P.S. [blog]," Retrieved November 24, 2007 from http://blog.pietrosperoni.it/2005/05/25/tag-clouds-metric/ 2005.

[50] C Shirky, "Power Laws, Weblogs, and Inequality. Clay Shirky's Writings About the Internet: Economics & Culture, Media & Community [blog]," Retrieved December 5, 2007 from http://www.shirky.com/writings/powerlaw_weblog.html 2003.

[51] G Smith, "Folksonomy: social classification. Atomiq / information Architecture [blog]," Retrieved December 15, 2007 from http://atomiq.org/archives/2004/08/folksonomy_social_classification.html 2004.

[52] J Udell, "Jon Udell: language evolution in del.icio.us," Retrieved December 15, 2005 from http://weblog.infoworld.com/udell/gems/delicious.html 2005.

[53] D. R., & Feinberg, J Millen, "Using Social Tagging to Improve Social Navigation.Workshop on the Social Navigation and Community-Based Adaptation Technologies In Conjunction with Adaptive Hypermedia and Adaptive Web-Based Systems (AH'06), Dublin, Ireland," Retrieved November 20, 2007 from http://www.sis.pitt.edu/~paws/SNC_BAT06/crc/millen.pdf. 2006.

[54] S Hayman, "Folksonomies and Tagging: New developments in social bookmarking. Ark Group Conference: Developing and Improving Classification Schemes, Rydges World Square," Sydney, Retrieved January 31, 2008 from http://www.educationau.edu.au/jahia/webdav/site/myjahiasite/shared/papers/arkhayman.pdf. 2007.

[55] X., Beaudoin, J. E., Bui, Y., & Desai, K Lin, "Exploring characteristics of social classification. Advances in Classification Research, Volume 17; Proceedings of the 17th ASIS&T Classification Research Workshop, Austin, Texas, USA. J. Furner & J. T. Tennis (Eds.).," Retrieved January 31, 2008 from http://dlist.sir.arizona.edu/1790/01/lin.pdf. 2006.

[56] M. E Kipp, "Complementary or Discrete Contexts in Online Indexing: A Comparison of User, Creator and Intermediary Keywords. Canadian Association for Information Science," Toronto, Ontario, Canada, Retrieved January 31, 2008 from http://dlist.sir.arizona.edu/1533/01/mkipp-caispaper.pdf 2006a.

Références bibliographiques

[57] M. E Kipp, "Exploring the context of user, creator and intermediate tagging. IA Summit 2006," Vancouver, BC, Retrieved January 31, 2008 from http://www.iasummit.org/2006/files/109_Presentation_Desc.pdf. 2006b.

[58] M. E Kipp, "Tagging Practices on Research Oriented Social Bookmarking Sites. Canadian Association for Information Science," Montreal, Quebec, Canada, Retrieved January 31, 2008 from http://dlist.sir.arizona.edu/2027/01/kipp%5F2007.pdf. 2007b.

[59] M. E Kipp, "Tagging for health information organisation and retrieval. ," North American Symposium on Knowledge Organization, Toronto, Ontario.J. T. Tennis (Ed.), Retrieved December 12, 2007 from http://dlist.sir.arizona.edu/1909. 2007a.

[60] B., & Hanser, C Berendt, "Tags are not metadata, but "just more content" - to some people.," Weblogs and Social Media, Boulder, Colorado, USA, Retrieved December 5, 2007 from http://www.icwsm.org/papers/2--Berendt-Hanser.pdf. 2007.

[61] T Smith, "Cataloging and You: Measuring the Efficacy of a Folksonomy for Subject Analysis. 18th Workshop of the American Society for Information Science and Technology Special Interest Group in Classification Research, Milwaukee, Wisconsin, USA. J. Lussky (Ed.).," USA, Retrieved January 31, 2008 from http://dlist.sir.arizona.edu/2061/01/Smith%5FUpdated.doc. 2007.

[62] P Mika, "Ontologies are us: A unified model of social networks and semantics," Journal of Web Semantics, 5(1), 5-15. 2007.

[63] I., Hamasaki, M., & Takeda, H Ohmukai, "A Proposal of Community-based Folksonomy with RDF Metadata. 4th International Semantic Web Conference," Galway, Ireland, Retrieved January 31, 2008 from http://www.ifi.uzh.ch/ddis/fileadmin/events/iswc2005ws/CameraReady/Ohmukai_F 2005.

[64] D Beckett, "Semantics Through the Tag. XTech 2006: Building Web 2.0," Amsterdam, The Netherlands, Retrieved January 31, 2008 from http://xtech06.usefulinc.com/schedule/paper/135. 2006.

[65] A Passant, "Using Ontologies to Strengthen Folksonomies and Enrich Information Retrieval in Weblogs: Theoretical background and corporate use-case," Weblogs and Social Media, Boulder, Colorado, USA., Retrieved December 5, 2007 from http://www.icwsm.org/papers/2--Passant.pdf. 2007.

[66] A., Tesconi, M., Ronzano, F., Rosella, M., & Minutoli, S Marchetti, "SemKey: A Semantic Collaborative Tagging System. Tagging and Metadata for Social Information Organization; a workshop at WWW2007, Banff, Alberta," canada, Retrieved December 5, 2007 from http://www.ibiblio.org/www_tagging/2007/paper_45.pdf 2007.

[67] A., Levialdi, S., & Malizia, A Dix, "Semantic Halo for Collaboration Tagging Systems. Workshop on the Social Navigation and Community-Based Adaptation Technologies In Conjunction with Adaptive Hypermedia and Adaptive Web-Based Systems (AH'06)," Dublin, Ireland, Retrieved January 31, 2008 from http://www.pitt.edu/~paws//SNC_BAT06/crc/malizia.pdf 2006.

Références bibliographiques

[68] Nicolas MORIN, "Bibliothèques, tags et folksonomies L'indexation des bibliothèques à l'ère sociale," Diplôme de conservateur de bibliothèque 2008.

[69] Anne Zemmour, "Indexation sociale et folksonomies : étude des principes d'organisation et de classement de l'information dans les bibliothèques 2.0," Mémoire de Master 1 Recherche 2011.

[70] M., et al. Fareh, "Merging ontology by semantic enrichment and combining similarity measures. International Journal of Metadata Semantics and Ontologies," 2013.

[71] [Online]. http://www.ista-ntic.com/v2/programmation/Java/Generalites/Generalites.htm

[72] [Online]. http://www.techno-science.net

[73] [Online]. http://www.eclipse.org

[74] James Allan, and Ben Carterette. Mark D Smucker, "A comparison of statistical significance tests for information retrieval evaluation. In Proceedings of the sixteenth ACM conference on Conference on information and knowledge management," pages 623–632 ACM, 2007.

[75] Jack Mills, and Michael Keen. Cyril Cleverdon, "Factors determining the rperformance of indexing systems volume 1. design.," Cranfield: College of Aeronautics 1966.

[76] MR. ABDELKRIM BOURAMOUL, "RECHERCHE D'INFORMATION CONTEXTUELLE ET SEMANTIQUE SUR LE WEB," Constantine, thèse 2011.

[77] Fuhr, "fuhr, N. models for reterieval with probabilistic indexing.information processing and management," 1989.

[78] 05-a M.Baziz, ""Indexation conceptuelle guidée par ontologie pour la recherche," Université Paul Sabatier de toulouse, 2005.

[79] Rijsbergen, "C.J.van rijsbergen information reterieval, 2nd butterworths," London, 1979.

[80] K. Järvelin and J. Kekäläinen., "Cumulated gain-based evaluation of ir techniques.," ACM Trans. Inf. Syst., 20(4):422–446, October 2002.

[81] ismail badache, "recherche d'information sociale exploitation des signaux sociaux pour ameliorer la recherche d'information," toulouse, 2016.

[82] [Online]. http://lyle.smu.edu/~tspell/jaws/index.html