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

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

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

Mémoire de Fin d'Etudes de MASTER ACADEMIQUE

Domaine : **Mathématiques et Informatique**
Filière : **Informatique**
Spécialité : **Conduite de projet Informatique**

Présenté par
Younes BEY
Nassim BEN RAMDANI
Malia MANSOURI

Thème

Calcul de Similarité entre annotations conceptuelles

Mémoire soutenu publiquement le 29/09/ 2016 devant le jury composé de :

Président : Mlle AIT ADDA Samia
Encadreur : Mlle ILTACHE Samia
Examinateur : M AMIROUCHE Nabil
Examinateur : Mme TAOURI Dalila

Remerciements :

« Aucun de nous ne s'est élevé à la seule force de son poignet.

Nous sommes arrivés parce que quelqu'un s'est baissé pour nous aider. »

Thurgood Marshall.

Avant tous, nous remercions Allah qui nous a donné la force, le courage et l'espoir nécessaire pour accomplir se travail.

Au terme de ce travail nous tenons à remercier notre promotrice Mlle ILTACHE pour son encadrement et ses précieux conseils.

Nous tenons à remercier également les membres de jury, pour avoir fait le plaisir d'accepter d'examiner ce travail.

Un immense merci à nos parents pour leur amour profond et leur soutien inconditionnel.

Nos remerciements aussi à tous ceux qui ont contribué à notre formation et qui nous ont soutenus dans nos études.

Dédicaces:

Je dédie ce travail à :

Ma mère, qui a oeuvré pour ma réussite, de par son amour, son soutien, tous les sacrifices consentis et ses précieux conseils, pour toute son assistance et sa présence dans ma vie, reçois à travers ce travail aussi modeste soit-il, l'expression de mes sentiments et de mon éternelle gratitude.

Mon père, qui peut être fier et trouver ici le résultat de longues années de sacrifices et de privations pour m'aider à avancer dans la vie. Puisse Dieu faire en sorte que ce travail porte son fruit ; Merci pour les valeurs nobles, l'éducation et le soutient permanent venu de toi.

Mes freres Khaled, Aissa, Zohir.

A mes chers grands parents.

A toute ma grande famille, mes oncles, mes tantes et mes cousins.

Tous mes compagnons de promotion, et à tous mes amis.

Younes

Dédicaces :

Je dédie ce travail à :

Ma mère, qui a oeuvré pour ma réussite, de par son amour, son soutien, tous les sacrifices consentis et ses précieux conseils, pour toute son assistance et sa présence dans ma vie, reçois à travers ce travail aussi modeste soit-il, l'expression de mes sentiments et de mon éternelle gratitude.

Mon père, qui peut être fier et trouver ici le résultat de longues années de sacrifices et de privations pour m'aider à avancer dans la vie. Puisse Dieu faire en sorte que ce travail porte son fruit ; Merci pour les valeurs nobles, l'éducation et le soutient permanent venu de toi.

Mon frère Yani et sœur Annela .

A mes chers grands parents.

A toute ma grande famille, mes oncles, mes tantes et mes cousins.

Tous mes compagnons de promotion, et à tous mes amis.

NASSIM

DÉDICACES :

Je dédie ce travail à :

Je dédie ce modeste travail et ma profonde gratitude à ma mére « Nacéra » et mon père « Meziane » pour l'éducation qu'ils m'ont prodigué; avec tous les moyens et au prix de toutes les sacrifices qu'ils ont consentis à mon égard, pour le sens du devoir qu'ils mon enseigné depuis mon enfance.

A ma sœur Lydia pour toute l'affection qu'elle m'a donnée et pour leur précieux encouragement. J'espère que la vie lui réserve le meilleur.

A mon adorable frère Chabane.

A ma chère sœur Massicélia et son mari Morad

A mes petites sœurs Sonia et Zazi.

A jedi CHABANE que dieu l'accueil dans son vaste paradis.
A jedi ESSAID que dieu le protège et le garde pour nous.
A mes très chères jida ZAZI et jida GHNIMA pour leurs chaleureuses bénédictions, Que dieu les protègent.
Mes tantes, oncles, cousins et cousines, et spécialement à khalti Ourida et sa petite famille qui m'ont soutenu au cours de ma formation.

A mes meilleurs amis :Didy, Kamilia, Mohamed, Nassim, Nassima, Rasha, Ronza, Yahia,Younes.

Tous mes compagnons de promotion .

A tous ceux qu'ils m'aiment.
A tous ceux que j'aime.

MALIA

Tables des matières

Tables des matières

Tables des matières

Liste des figures

Listes des figures

Liste des tableaux

Liste des tableaux

Introduction générale

Introduction générale

Avec l'augmentation rapide du volume d'information stocké sous format numérique, et l'avènement du Web, la quantité d'informations disponible ne cesse de croître au cours de ces dernières années, il est devenu alors très difficile de trouver une information ou un document qui répond à un besoin de l'utilisateur. Jusqu'ici on dispose d'un grand volume d'information, mais sans aucune maîtrise de contenu, le résultat est que l'utilisateur perd beaucoup de son temps à examiner un grand nombre de document en cherchant ce qui lui convient, Il a fallu donc envisager le développement des outils automatiques qui permettent de conserver, chercher et classer ces informations, et d'assurer une utilisation ciblé et efficaces de ces données.
Notre travail traite l'utilisation des mesures de similarité sémantique pour exprimer la ressemblance entre les documents textuelle.

Nous avons décomposer notre mémoire en quatre chapitres. Le premier chapitre vise à définir le processus d'indexation avec ces différents étapes et le langage utilisé qui mène au calcul de la similarité, on présentera aussi l'algorithme de porter, qui est utilisé pour la normalisation des mots. Ainsi, les modèles de représentation des documents textuels.

Nous enchainerons dans le deuxième chapitre le problème général de l'indexation classique, et on introduira comme solution l'indexation sémantique basé sur le sens des mots, et pour trouver le sens correcte de mots différentes approches de désambiguïsation sont exploités, basées sur l'utilisation des ressources lexicales et sémantiques.

Le troisième chapitre présente l'utilisation de la similarité dans les différents domaines : la recherche d'information, la détection plagiat et dans le domaine de classification des documents .ainsi, les approches utiliser pour mesurer la similarité, qui est notre objectif principal de notre travail.

Enfin, le dernier chapitre expose la description des approches implémentées ainsi que les résultats obtenus.

Chapitre I : Indexation des documents

Introduction :

Le but de l'indexation est de créer une représentation permettant de repérer et retrouver facilement l'information dans un ensemble de documents. [Lancaster, 1998] donne cette définition : "Le but principal de l'indexation (et du résumé automatique) est de construire des représentations d'éléments publiés sous une forme adaptée pour le stockage dans tout type de base de données".

On utilise cette indexation, le plus souvent, pour les systèmes de recherche d'informations. Mais, elle peut également servir à comparer et classifier des documents, proposer des mots clés, faire une synthèse automatique de documents, calculer des co-occurrences de termes… Tout index de document perd une partie de l'information initiale.

I. Langage d'indexation :

Lors de la conception d'un SRI, la question cruciale du langage d'indexation ou vocabulaire d'indexation [Cleveland &al., 2000] arrive en premier. Ce langage peut être libre ou contrôlé. [Abichahine., 2011].

I.1. Langage libre :

Construit à partir des termes en langue naturelle, souvent issus du texte original, et permettant de décrire son contenu [Harter,1986] (on parle d'indexation par extraction ou extraction indexing en anglais). On trouve ce type d'indexation par vocabulaire libre dans les moteurs de recherche de type GOOGLE.

I.2. Langage controlé :

Construit à partir des termes extraits d'un thésaurus. Le thésaurus est une liste de descripteurs (mots-clés) normalisés et reliés entre eux par des relations sémantiques [Boucham, 2009].

Ces relations sont au nombre de trois [Roussey&al.,2001]:

- La relation d'équivalence regroupe les termes jugés équivalents (synonymes ou termes très proches sémantiquement).
- La relation hiérarchique construit une hiérarchie entre les termes d'indexation, du général au particulier ou d'un tout à ses parties.
- La relation d'association lie des termes d'indexation ayant des connotations (basées sur la cooccurrence des termes).

Selon [Boucham., 2009] l'organisation du thésaurus permet de trouver le terme d'indexation le plus approprié pour représenter un concept. Par exemple, l'utilisateur d'un système de recherche d'information utilise un terme de son vocabulaire comme entrée dans le thésaurus et, en suivant différentes relations, trouve le terme d'indexation reconnu par le système pour composer sa requête.

II. Les modes d'indexation :

L'indexation peut se faire de 3 manières différentes : manuellement (faite par un humain), de manière semi-automatique (par exemple créée par un humain assisté d'un programme proposant des termes), ou de manière automatique (créée par un programme informatique) **[Kompaoré., 2008].**

I.1. L'indexation manuelle : l'indexation est réalisée par un spécialiste du domaine correspondant ou un documentaliste qui analyse le document et choisit une liste de descripteurs. Il assure une meilleure précision, Cependant il représente des inconvénients d'être couteuse et subjective (le même document peut être indexé différemment par différent personnes).

I.2. L'indexation semi-automatique ou indexation supervisée : cette méthode est un assemblage des deux autres méthodes, A la fin de l'indexation automatique la liste des termes issus est exposé au documentaliste pour choisir les descripteurs final en éliminant quelques termes et validant d'autres.

I.3. L'indexation automatique : dans ce cas, chaque document est analysé à l'aide d'un processus entièrement automatisé.Cette méthode fournit toujours le même indexe pour le même document, elle se réalise par différent étapes qui sont **:** La tokenisation, L'élimination des mots vides, La normalisation, la sélection et la pondération.

III. Segmentation de textes :

Un programme d'indexation, s'il travaille sur de grands documents, pourra tenir compte des différentes unités d'indexation (unité linguistique) qui sont la phrase, le paragraphe, ou le document dans son ensemble. Ce qui implique l'utilisation d'un programme de segmentation. Le plus simple étant de reconnaître une phrase comme étant une suite de mots suivie d'un point. Lorsque l'on travaille sur des textes ayant un format défini, on peut parfois extraire la notion de phrase ou de paragraphe en analysant le format (SGML, XML, dans une moindre mesure HTML). Une autre solution consiste à utiliser des étiqueteurs syntaxiques qui permettent, notamment, d'identifier les paragraphes et les phrases d'un texte.

L'indexation de textes est le plus souvent une étape préalable d'un système plus complet comme la recherche d'informations, l'attribution de mots-clés, la similarité de documents ou la synthèse automatique.

IV. Les étapes de l'indexation automatique :

L'indexation se décompose en trois phases : schématisées dans la figure 1.

- L'extraction des termes du document.

- La sélection des termes discriminatifs pour un document.

- La pondération des termes.

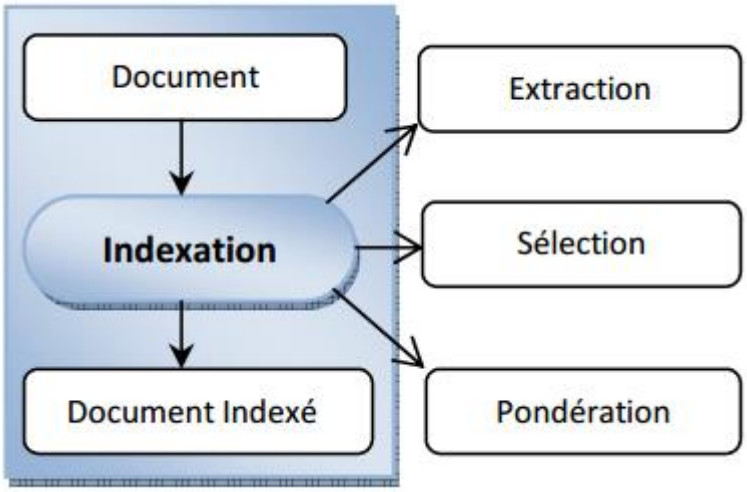


Figure 1 : Indexation d'un document [Tambellini, 2007]

IV.1. L'extraction des termes du document :

Cette phase est composée de trois étapes, ces étapes sont :

IV.1.1. *L'analyse lexical (Tokenisation) :* cette étape consiste à découper tout le texte de document (ou requête) en un ensemble de terme (lexème), et éliminer les espaces, les ponctuations, la casse, et la mise en page.

IV.1.2. *L'élimination des mots vides :* cette étape permet d'éliminer tout les mots vides (pronoms personnels, conjonctions, prépositions,...) qui apparaissent d'une manière uniforme dans les documents et ils sont non utiles pour l'indexation, et pour cela on utilise des stoplist ou des anti-dictionnaires.

Cependant, il ne faut pas oublier de tenir compte de certains mots vides qui auraient pour homographes des mots significatifs comme par exemple la conjonction de coordination « or » qui peut utilisé pour référer a «l'or » qui est un métal précieux.

IV.1.3. *La normalisation :* ce processus permet de trouver la forme normale d'un mot donné, plusieurs stratégies de normalisation sont utilisées On peut citer l'algorithme de Porter, qui a pour but de trouver le radical d'un mot anglais en supprimant sa terminaison.

Ce traitement repose sur deux principales procédures, la racinisation et la lemmatisation.

- *la racinisation :* le stemming(en anglais) elle est utilisée pour éliminer les variations morphologiques ou les variations orthographiques des mots clés.

Elle consiste à supprimer les préfixes et les suffixes des termes, et représenter uniquement le radical de mot donné.

- *la lemmatisation* : ce processus permet de trouver la forme canonique d'un ensemble de mot de la même catégorie grammaticale (par exemple ferm pour fermer, fermable, fermé, fermeture …).

IV.1.4. L'algorithme de porter :

L'algorithme de Porter est un algorithme de normalisation des mots. Il permet de supprimer les affixes des mots pour obtenir une forme canonique du mot. Cet algorithme est utilisé pour la langue anglaise, mais son efficacité est limitée pour la langue française où les flexions sont plus importants et plus diverses. Il reste toutefois un algorithme fondamental couramment enseigné en TALN[1].

1. *Le principe de la normalisation :*

Lorsque l'on emploie un terme dans une langue flexionnelle ou agglutinante, celui-ci subit des flexions. Une flexion est une modification morphologique d'un terme afin de marquer la position grammaticale, le temps de conjugaison, ... Par exemple, le verbe manger se fléchit en "mangeons" lorsqu'il est placé dans une phrase au présent et a pour sujet une première personne du pluriel. Le mot cheval se fléchit en chevaux au pluriel, ...

L'objectif de la normalisation (stemmatisation ou encore lemmatisation) est d'effectuer l'opération inverse, ie retrouver la forme canonique commune à des mots fléchis.

Si l'on prend la phrase :

Les chevaux courent dans les prairies.

La normalisation des mots la composant donne :

Le cheval courir dans le prairie.

2. *Le découpage en pseudo-syllables :*

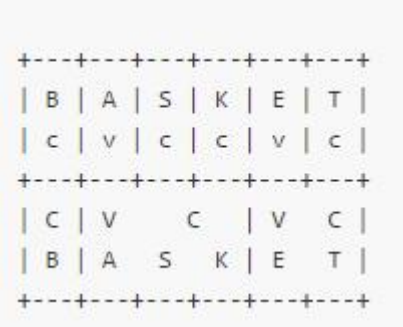
L'algorithme de Porter travaille sur les mots, mots composés à leur plus bas niveau de lettres. On peut partitionner l'ensemble des lettres de l'alphabet latin en deux classes :

- les voyelles (notées **v**) : A, E, I, O, U et Y quand il est précédé d'une consonne ;
- les consonnes (notées **c**) : toutes les autres lettres et Y quand il est précédé d'une voyelle ;

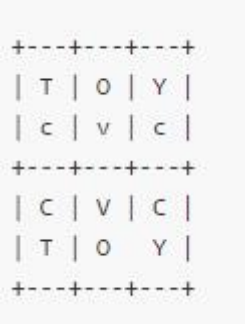
D'après cette classification des lettres, chaque mot peut s'écrire [C]VC...[V]avec V une séquence d'au moins une voyelle et C une séquence d'au moins une consonne. Les couples VC correspondent alors aux pseudo-syllables du mot. Le nombre de pseudo-syllables du mot est appelée la **mesure du mot** et se note **m**.

[1]:https://tartarus.org/martin/PorterStemmer/

Si l'on prend le mot basket, le découpage présenté ci-dessous révèle deux pseudo-syllables (ASK et ET), d'où la mesure du mot basket est de deux (m=2).



Si l'on prend le mot toy, le découpage présenté ci-dessous révèle une seule pseudo-syllable, la mesure du mot toy est donc d'une (m=1). À noter ici que la lettre Y étant précédée de la voyelle O, elle est considérée comme une consonne.



3. ***Les règles de transformation :***

Les règles de transformation morphologiques introduites par Porter ont la syntaxe suivante : **(<condition>) S1 -> S2**, avec :

- **S1** est le suffixe que l'on retire au mot afin d'obtenir le radical ;
- **S2** est le suffixe que l'on ajoute au radical si la condition est remplie ;
- **condition** est une condition que le radical doit vérifier pour que la règle soit appliquée.

La condition s'applique sur le mot considéré privé de S1, partie du mot qu'on appellera radical par la suite, Il y a un certain nombre de notations spécifiques à certaines conditions :

- *S signifie que le radical se termine par la lettre S ;
- *v* signifie que le radical contient une voyelle ;
- *d signifie que le radical se termine par deux consonnes ;
- *o signifie que le radical se termine par la séquence cvc (une consonne, une voyelle et une consonne) et que la dernière consonne n'est ni W, ni X, ni Y.

La dernière notation peut sembler quelque peu farfelue, mais il ne faut pas oublier que l'algorithme a été mis au point pour l'anglais.

Si le radical du mot est BREW, la condition (*W) sera remplie, mais pas la condition (*o) car bien que REW respecte la séquence cvc, W fait partie ne peut pas être la dernière lettre du radical.

4. ***Application de l'algorithme :***

L'algorithme de Porter permet donc d'appliquer des règles définies dans une syntaxe particulière sur des mots fléchis. L'application des dites règles permet de réaliser des transformations morphologiques afin d'obtenir une version normalisée à partir d'une version fléchie.

Quelques exemples bien choisis étant plus parlant qu'un beau discours, considérons les règles suivantes :

- (m>0) EED -> EE ; soit EED le suffixe à retirer pour obtenir le radical, et EE le suffixe à ajouter au radical si la condition m>0 est remplie (m est la mesure du mot, soit son nombre de pseudo-syllables, cf plus haut) ;
- (*v*) ED -> ; soit ED le suffixe à retirer pour obtenir le radical, et l'on n'ajoute aucun suffixe au dit radical si la condition *v* est remplie (*v* signifie que le radical contient une voyelle cf plus haut) ;
- (*v*) ING -> ; à vous de deviner ...

Prenons le mot agreed, en supprimant le suffixe EED, on obtient le radical agr. Ce radical contient une pseudo-syllable et rempli donc la condition m>0, on y ajoute donc le suffixe EE pour obtenir agree, la forme normalisée de agreed.

Prenons maintenant le mot sing, on tente d'y appliquer la troisième règle, puisque les suffixes des deux premières ne correspondent pas. On obtient donc le radical s. Ce dernier ne contient aucune voyelle et ne remplit donc pas la condition *v*. On conserve donc la forme sing comme forme normalisée étant donné qu'il n'y a aucune autre règle qui peut s'y appliquer.

Prenons finalement le mot spending, en supprimant le suffixe ING on obtient le radical spend qui contient effectivement une voyelle. La condition *v* étant remplie, on applique la règle, et on ajoute donc le suffixe S2 au radical pour obtenir la forme normalisée. Étant donnée qu'il n'y a pas de suffixe S2, le radical correspond à la forme normalisée : spend.

IV.2. La sélection des termes discriminatifs pour un document : Une fois les termes candidats à l'indexation extraits et normalisés, il reste à faire un choix sur ceux qui seront effectivement retenus pour l'indexation. Deux alternatives sont possibles :

- Retenir tous les termes identifiés comme terme d'index, on parle alors d'indexation **full-texte.**
- Ne retenir des termes candidats que ceux intéressants pour l'indexation, on parle alors de l'indexation **sélective**.

IV.3. La pondération des termes : 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, [Roberston., 1976] 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 [Robertson., 76], [Singhal., 1997] et [Sparck Jones., 1979]. On y distingue :

IV.3.1. *Tf* (*term frequency*) : 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 :

1. Tf : utilisation brute,
2. $0.5 + 0.5 \quad \frac{Tf}{Max(Tf)}$

IV.3.2. *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 :

1. $Idf = \log\left(\frac{N}{df}\right)$,
2. $Idf = \log\left(\frac{N - df}{df}\right)$.

Où *df* est la proportion de documents contenant le terme et N le nombre total de documents dans la collection.

La fonction de pondération de la forme ***Tf.Idf*** consiste à multiplier les deux mesures Tf et Idf. Une formule largement utilisée est la suivante:

$$Tf.Idf = \left(0.5 + 0.5\frac{Tf}{Max(Tf)}\right) * \log\left(\frac{N}{df}\right)$$

Une normalisation de la mesure du *Tf.Idf* par rapport à la longueur des documents a été proposée par [Singhal, 1995].

$$Tf.Idf = \frac{Tf + \log\left(\frac{N - df + 0.5}{df + 0.5}\right)}{2.\left(0.25 + 0.75.\frac{dl}{\Delta d}\right)}$$

dl est la longueur du document en nombre de termes et Δd la longueur moyenne des documents de la collection.

En effet, 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 [DeClaris., 1994].

V. La représentation des documents :

Il existe dans la littérature de nombreux modèles de représentation de documents textuels. Nous pouvons citer, par exemple, l'utilisation de vecteurs dont les composantes représentent des termes [Sal 1973, Sal 1989], des matrices de distribution de termes ou de relations entre termes [Bes&al 2001, Lan&Lit 1991].

L'objectif de la suite de cette section est d'étudier les différents modèles de représentation de textes largement utilisés.

V.1. *Représentations vectorielles :*

Les modèles vectoriels sont largement utilisés pour la représentation de textes. Le modèle vectoriel standard et le modèle Latent Semantic Indexing sont les plus utilisés et implémentés. Nous détaillons ces derniers dans la suite de cette section.

V.1.1. *Le modèle standard « sacs de mots » :*

Dans le cadre du modèle vectoriel standard (MS), les textes sont considérés comme des « sacs de mots » [Sal 1971a], [Sal 1971b], [Sal&McG 1983]. L'idée principale est de transformer les différents documents d'une base documentaire en vecteurs où chacun des éléments d'un vecteur de texte représente des unités textuelles ou tout simplement des mots appelés aussi « termes d'indexation ».Plusieurs travaux utilisent les mots comme termes d'indexation [Dum&al 1998],[Aas&Eik 1999], [Apt&al 1994], [Lew 1992b]. Un mot est considéré comme étant une suite de caractères encadrés par des caractères de ponctuation ou appartenant à un dictionnaire spécifique. Des outils, basés sur des approches linguistiques, statistiques ou mixtes sont utilisés pour l'identification des différentes unités textuelles.

Dans le modèle vectoriel standard, les composantes d'un vecteur représentant un texte sont fonction de l'occurrence des mots dans le texte. Ce modèle a été initialement introduit par Gérard Salton [Sal&Les 1965], [Sal 1971a] dans l'objectif d'implémenter un système de recherche d'informations. L'implémentation la plus connu de ce modèle est le système de recherche documentaire SMART.L'évolution de ce système est décrite dans [Sal 1991]. Dans ce modèle les composantes des vecteurs représentent des termes considérés comme les plus discriminants. Dans le cadre du modèle vectoriel standard, ces termes sont sélectionnés en fonction de leurs fréquences d'apparition dans les documents et en fonction du nombre de documents contenant ces termes.

Soit C une collection de documents textuels, Di un document de C, soit t le nombre de termes d'indexation et $T = \{T_1,\ldots, T_j,\ldots,T_t\}$ l'ensemble de ces derniers. Dans le modèle vectoriel standard, le document Di est représenté par un vecteur Vi. La collection de textes peut être ainsi représentée par une matrice dont les colonnes représentent les termes d'indexation et les lignes représentent les documents de cette collection.

$$Vi = (W1,\ldots, Wj,\ldots Wt),$$

Où Wj est le poids d'un terme Tj dans le document Di et j = 1..t.

Le poids donné à un terme d'indexation dans un document est calculé en fonction de la fréquence d'occurrence du terme TF (Term Frequency) dans le document et du nombre de documents contenant le terme IDF (Inverse Document Frequency). Les calculs des poids et les pondérations accordés à un document D ont fait l'objet de nombreuses études [Sin 1997], [Lee 1995], [Buc&al 1992], [Sal&Buc 1988].

V.1.2. ***Représentations conceptuelles et basées thésaurus :***

Ils existent des approches utilisant les concepts pour la représentation des textes tel que Word Category Map ou des modèles issus de la méthode LSI [Koh&al 2000], [Dee&al 1990], [Lio&al 2004]. Ces méthodes dépendent de la distribution des probabilités des mots au sein du jeu d'apprentissage, en conséquence les données du jeu ont une influence sur la génération de l'espace des concepts. J. Chauché [Cha 1990] a proposé un nouveau modèle vectoriel de représentation de textes. Au lieu de définir un espace vectoriel dont chaque dimension représente un terme d'indexation, l'ensemble des termes est projeté sur un ensemble fini de concepts extraits d'un thesaurus. Cette représentation permet une factorisation des termes par regroupement de leurs champs sémantiques. Par exemple, deux synonymes partageront un ensemble de mêmes concepts. L'auteur utilise, pour des documents français, un thésaurus composé de 873 concepts hiérarchisés en 4 niveaux. Rappelons qu'un thésaurus permet uniquement d'explorer, à partir d'un concept, les mots qui s'y rattachent et inversement.

Par exemple, le mot « interview » défini par les concepts 419 (question), 583 (compagnie) et 726 (communication), 749 (conversation) et 766 (presse) du thésaurus, sera représenté par un vecteur de dimension 873 dont toutes les composantes sont nulles sauf celles associées aux concepts 419, 583, 726, 749 et 766 qui seront identiques (C.f. Figure 2). Le thésaurus est donc défini comme un ensemble de couples de L × R873 avec L l'ensemble des lemmes du thésaurus. Les dimensions de l'espace vectoriel ne sont pas associées à des termes d'indexation mais à des concepts.

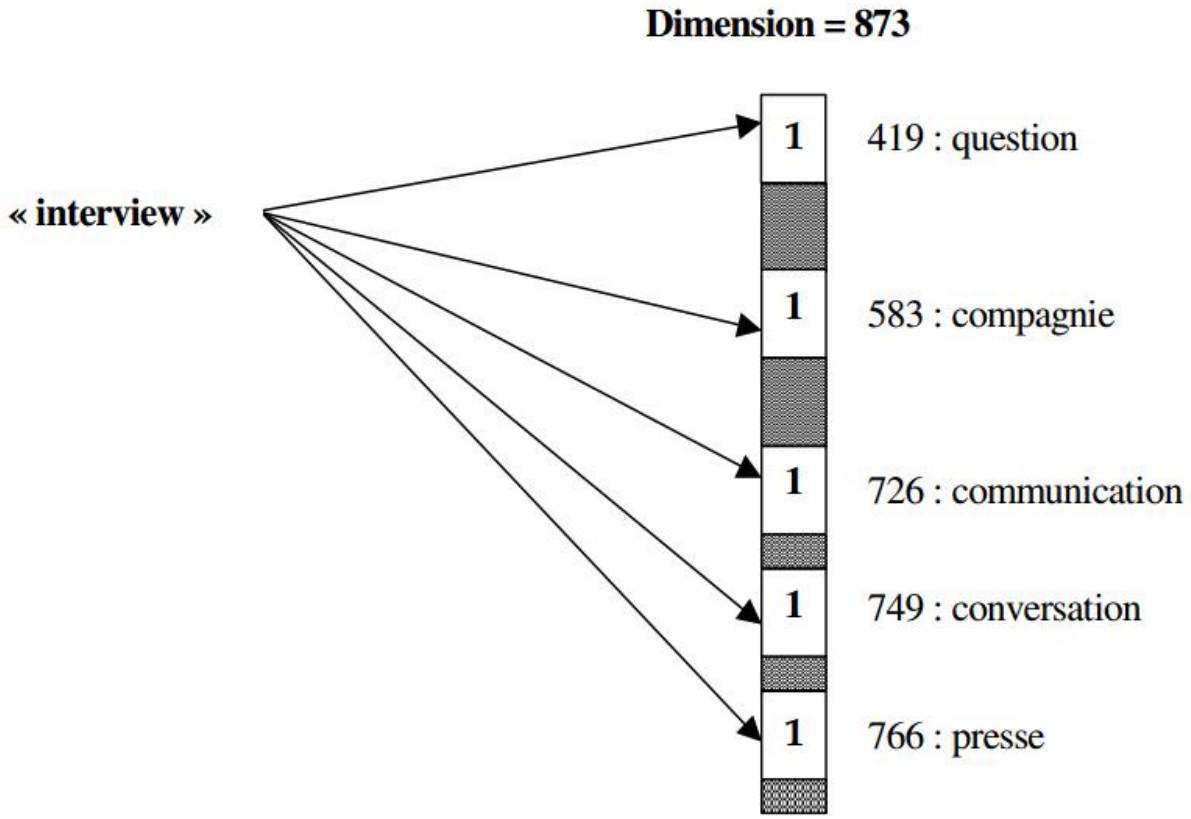


Figure 2 : Exemple de représentation conceptuelle du mot « interview »

Après l'extraction de l'ensemble des lemmes d'un texte, une association est réalisée entre les lemmes et le vecteur qui leur est associé au sein du thésaurus. Ensuite, pour chaque texte un vecteur conceptuel est calculé en fonction de la moyenne normalisée des lemmes qu'il contient [Jai&al 2005].

V.2. Représentations basées sur les associations de termes :

Nous avons présenté dans la section précédente des modèles de représentations vectorielles de textes. Ces modèles exploitent essentiellement la structure explicite des documents (les mots). Certains d'entre eux tentent de prendre en considération les dépendances qui peuvent exister entre les mots par des modèles statistiques ou des transformations stochastiques en partant des informations sur l'occurrence des mots dans les documents (TF) ou en documents (IDF). L'origine de ces approches est issue de la Sémantique Distributionnelle (SD) qui est fondée sur l'hypothèse suivante : « La sémantique des éléments textuelles est déterminée par leurs distributions dans les textes ». En effet, elle suppose l'existence d'une forte corrélation entre les caractéristiques distributionnelles observables des mots et leurs sens. Ainsi, la sémantique d'un mot est reliée à l'ensemble des contextes dans lesquels apparaît ce dernier [Har 1988], [Har&al 1989]. Notons que la sémantique distributionnelle hérite de la théorie de Firth « The basic assumption of the theory of analysis by levels is that any text can be regarded as a constituent of a context of situation, ……, You shall know a word by the company it keeps » [Fir 1957] qui peut se résumer par « le sens d'un mot peut être donné par ses voisins ».

V.2.1. ***Association de termes :***

En raison des difficultés de représentation des connaissances textuelles par des modèles symboliques structurés tel que les « frames » et les modèles logiques, des méthodes plus adaptées aux applications exploitant des collections de documents sont nécessaires. Ces méthodes ont suscité un intérêt particulier ces dernières années. Nous nous intéressons particulièrement dans la suite de cette section aux modèles de représentation basés sur la notion d'association de termes. D'une manière générale, dans la littérature, la cooccurrence (appelée également association ou co-citation) de termes est définie de la manière suivante :

Définition 1 (Cooccurrence de termes)

Soient deux termes T1 et T2. Une cooccurrence entre les termes T1 et T2 est définie comme étant l'apparition commune des deux termes dans un même contexte. Elle correspond à deux termes qui apparaissent ensemble dans un même segment de texte. En fonction des modèles un segment de texte peut être une fenêtre de mots, une phrase ou un paragraphe.

Conclusion :

Dans ce chapitre, nous avons présenté les bases de l'indexation, à savoir, les langages d'indexation, les modes d'indexation et les étapes d'indexation

Nous avons aussi présenté dans ce chapitre les modèles de représentation de collections de documents textuels. Parmi ces derniers, les plus utilisés sont les modèles vectoriels, le modèle vectoriel standard (MS) via sa présentation en « sac de mots ». Le MS a été enrichi par le modèle LSI qui procède à des transformations sur la matrice M (documents × unités linguistiques).

L'indexation classique est basée que sur la notion des mots clés, qui se fait juste par la correspondance lexicale des termes, qui n'est pas généralement suffisante pour obtenir des informations consolidables. Cela est dû à l'ambigüité qui est un caractère inhérent aux noms dans la langue naturelle. Par ailleurs, nous envisageons dans le chapitre suivant des techniques de désambiguïsation du sens.

Chapitre II : Indexation sémantique

Introduction :

Les modèles classiques de la RI présenté dans le chapitre précédent se basent sur l'hypothèse qu'il y a une correspondance stricte entre les mots et les sens, alors qu'un mot peut représenter plusieurs sens et un sens peut être représenté par plusieurs mots. En partant de cette hypothèse, les chercheurs ont proposé d'utiliser une indexation autre que l'indexation classique, appelée indexation sémantique.

Dans ce chapitre, nous abordons la problématique de l'indexation classique basée mots-clés, en introduisant comme solution à ce problème l'indexation sémantique basée sur la désambiguïsation, ainsi que les approches mises en œuvre.

I. Problématique :

En indexation classique, la représentation du contenu d'un document, ou d'une requête utilisateur, par des mots-clés est généralement imprécise. Cette imprécision est causée par deux principaux problèmes [Boubekeur et al, 2010a], [Boubekeur et al., 2010b]: l'ambiguïté des mots de la langue et leur disparité lors de la recherche.

- L'ambiguïté des mots, ou ambiguïté lexicale, se rapporte à des mots lexicalement identiques et portant des sens différents. Elle se divise en deux catégories : l'ambiguïté syntaxique et l'ambiguïté sémantique.
  - L'ambiguïté syntaxique se rapporte à des différences dans la catégorie syntaxique. Par exemple, « play » peut apparaître 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 [Krovetz, 1997].
- La disparité des mots (wordmismatch) : se réfère à des mots lexicalement différents mais portant un même sens. En effet, des documents pourtant pertinents, ne partageant aucun mot avec la requête, ne seront pas restitués par le processus de recherche.

Pour surmonter ces problèmes, la RI sémantique est apparue. Son but est d'incorporer l'information sémantique dans le processus de la RI. On distingue deux grandes approches : l'ïndexation sémantique et l'ïndexation conceptuelle. On parle de l'indexation sémantique quand il s'agit d'utiliser le sens des mots (mot-sens ou word-sens) pour indexer les documents. L'indexation conceptuelle peut être vue comme une généralisation de l'indexation sémantique, dans la mesure où les concepts aussi véhiculent des sens [BAZIZ, 2005].

II. L'indexation sémantique (Sense Based Indexing) :

Est une approche d'indexation qui consiste à représenter les documents et les requêtes par les sens des mots (ou concepts) plutôt que par les mots eux-mêmes.Elle utilise des techniques de désambigüisation des mots, WSD (Word SenseDisambiguation), Elle fait associer aux mots extraits, des mots du contexte qui aident à déterminer leur sens.

Il s'agit en effet de déterminer d'abord le sens correct de chaque mot dans le document (ou la requête), ensuite le (la) représenter.

Par ailleurs, la désambiguïsation sémantique doit de ce fait permettre la sélection de l'information la plus riche et la plus précise, ce qui implique l'utilisation d'une ressource lexicale et sémantiques bien structurée.

III. Les ressources sémantiques externes :

III.1. Dictionnaire :

Un dictionnaire est un ouvrage de référence contenant l'ensemble des mots d'une langue ou d'un domaine d'activité généralement présentés par ordre alphabétique et fournissant pour chacun une définition, une explication ou une correspondance (synonyme, antonyme, cooccurrence, traduction, étymologie) [**https://fr.wikipedia.org/wiki/Dictionnaire].**

Dans le domaine de Recherche d'information nous parlons de dictionnaire automatisé ou informatisé, qui est un dictionnaire sous forme électronique qui peut être interrogé via une application.

III.2. Réseaux sémantique :

En recherche d'information, réseau sémantique (semantic network ou net) est représenté par des nœuds qui sont relier par des arcs, où les nœuds sont des concepts et les arcs représentent différent types de relations entre concepts" **[Lee et al, 1993].**
[Quillian, 1968] définit donc un réseau sémantique comme étant "un format de représentation permettant de mémoriser le sens des mots, pour rendre possible leur utilisation à la manière de l'être humain".
Si les réseaux sémantiques s'expriment par une représentation simple, il n'en subsiste pas moins quelques défauts, dont deux majeurs. Le premier est que les modèles développés sont rarement appliqués à des problèmes concrets par manque d'explications concernant leur utilisation. Ces modèles sont très souvent développés pour étudier leur pouvoir de représentation par rapport à des formalismes existants. Il existe plusieurs modèles bien documentés, dont celui des graphes conceptuels. Cette théorie a été développée en 1984 par John Sowa [Sowa, 1984]. Le second point est lié aux nombreux problèmes qui surviennent lors de la définition des nœuds (concepts, relations…) et des liens (difficultés d'interprétation…) [Brachman, 1983], [Woods, 1975]. Il est donc indispensable d'adopter une méthodologie de travail qui tienne compte de la réalité qu'exige un domaine aussi pratique que celui de la recherche d'information.
Pour avoir une idée sur la façon dont les réseaux sémantiques sont construits, prenons comme exemple un terme commun mais évocateur "oiseau". Écrivons-le au milieu d'une page blanche. Pensons à quelques termes liés au terme "oiseau" comme "voler" et "pigeon".
Ecrivons ces termes tout autour de "oiseau" et lions chacun d'eux par une ligne vers "oiseau". Donnons à chaque ligne une étiquette qui décrit la relation entre les deux termes – par exemple, la ligne reliant "oiseau" "voler" pourrait être étiquetée "peut". Continuons à l'extérieur à rajouter des termes liés au terme "oiseau", des termes reliés au terme "voler", etc. La figure suivante est un réseau sémantique de l'exemple précédent :

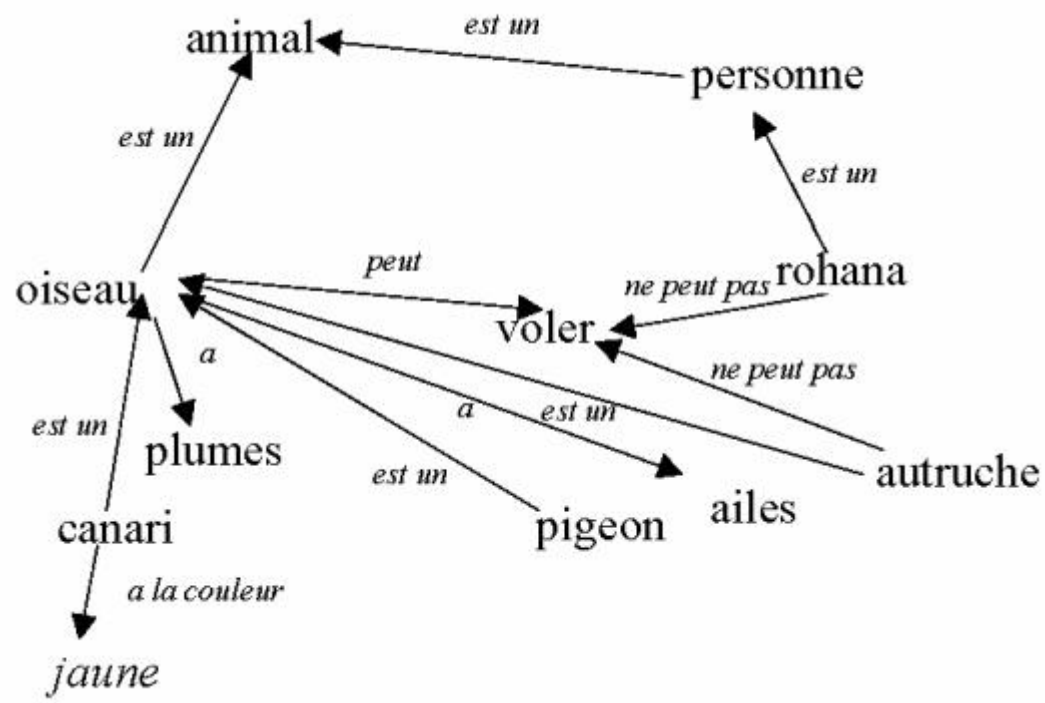


Figure 3 : Exemple de Représentation de Réseau Sémantique

III.3. Taxonomie :

La taxinomie est un type d'un réseau sémantique ou l'unique relation est un lien hiérarchique correspondent à des liens de (Spécialisation / Généralisation), souvent cette relation hiérarchique utilise liens comme « est un » (en anglais, «is a »).

Voici un exemple classique de taxinomie : Relation de division
–Monde
∟Afrique
∟Europe
∟France
∟Italie
Cet exemple montre une décomposition hiérarchique ou division du monde en continents et chaque continent en pays.

III.4. Thésaurus :

Un thésaurus constitue un dictionnaire hiérarchisé des vocabulaires contrôlés. Ce vocabulaire est constitué d'un ensemble de terme normalisés, liés entre eux par des relations sémantique (hyperonymie, hyponymie, synonymie, antonymie,…etc.) servant l'indexation des documents.

Parmi les thésaurus existant on cite : le thésaurus médical Mesh(MedicalSubjectHeading), et le méta-thésaurus médical UMLS (UnifiedMedicalLanguage System).

III.4.1. Thésaurus MeSH : : MeSH (Médical SubjectHeading), est un thesaurus contenant un vocabulaire contrôlé du domaine médical et un ensemble riche de relations liant les

différents termes. Il est utilisé pour indexer des articles et ouvrages traitant du domaine médical. MeSH comprend essentiellement des termes qui désignent les concepts biomédicaux, et les relations.

III.4.1.1. Terme : Un terme est un mot ou un ensemble de mots exprimant une notion particulière.

III.4.1.2. Concept : Un concept comprend un ou plusieurs termes synonymes et porte le nom d'un de ces termes, dit *terme préféré.*

III.4.1.3. Relation : Dans MeSH il existe deux types de relations entre les concepts : les relations hiérarchiques et les relations associatives (associé à). La hiérarchie dans MeSH est représentée par un code reflétant l'arborescence auquel le concept appartient et peut véhiculer plusieurs sens :

1. relation "*est une partie de*" (méronymie), par exemple le concept "*doigt*" (A01.378.800.667.430) est une partie de "*main*" (A01.378.800.667).

2. relation "*est un type de*" (hyponymie), par exemple le concept *pouce* (A01.378.800.667.430.705) est un type de "*doigt*" (A01.378.800.667.430).

3. relation "*est sémantiquement proche de*" (aboutness), par exemple le concept "*sécurité*" (G03.850.110.060.075) est sémantiquement proche de "*accidents*" (G03.850.110).

III.4.2. Méta-thésaurus UMLS :

Proposée et maintenue par la NLM(NationalLibraryofMedicine),l'UMLS est la ressource terminologique la plus large actuellement disponible pour la médecine. Elle est le résultat de la fusion de plus d'une centaine de thésaurus (dont MeSH, CIM-10 et SNOMED) de différentes langues, dont elle préserve les réseaux de relations entre termes. De ce fait, l'UMLS est qualifié de méta thésaurus. L'UMLS organise les termes autour de concepts ; il y a ainsi plus de 700 000 concepts, auxquels sont rattachés des ensembles de termes qui les désignent. Enfin, à chaque terme est associé un ensemble de variantes graphiques (les chaînes, issues par exemple de l'emploi différent des casses, des ponctuations, etc).

III.5. Les ontologies :

Une ontologie est un ensemble structuré de concepts, qui permet de modéliser un ensemble de connaissances dans un domaine donné. Les ontologies permettent, d'une part de décrire les connaissances d'un domaine spécifique et d'autre part de représenter des relations complexes entre les concepts.

Plusieurs éléments ou composants constituent une ontologie. Ceux qui reviennent le plus dans la littérature sont, (1) *les concepts* (souvent représentés par des termes), (2) *les relations* entre ces concepts (telles la relation *sous-classe-de* ou encore *partie-de*), (3) *les fonctions*, qui sont des cas particuliers de relations dans lesquelles le nième élément de la relation est défini de manière unique à partir des n-1 premiers, (4) *les axiomes*, utilisés pour

structurer des "phrases" qui sont toujours vraies et (5) *les instances* : elles sont utilisées pour représenter des éléments [Baziz M., 2005].

III.5.1. Les types de l'ontologie :

Nous listons ci-dessous les différents types d'ontologies les plus utilisées:

III.5.1.1. Ontologies de représentation : N'appartiennent à aucun domaine, mais définissent et organisent les primitives de la théorie logique pour permettre la représentation des ontologies. L'exemple le plus représentatif de ce genre d'ontologie est la Frame Ontologie13, qui définit d'une manière formelle, les primitives de représentation (classes, sous classes, attributs, valeurs, relations et axiomes) dans un environnement implémentant les langages de Frame **[Van Heijst& Schreiber &Wielinga, 1997].**

III.5.1.2. Ontologies génériques : aussi appelée Ontologie de haut niveau elles décrivent des concepts généraux, indépendants d'un domaine ou d'un problème particulier. Elles permettent par exemple de formaliser les aspects temporels ou spatiaux des objets du monde réel. Cyc14 est un exemple d'une ontologie générique portant sur des concepts de haut niveau. Ces dernières décrivent des notions générales comme les notions d'objet, de propriété, d'état, de valeur, de moment, d'évènement, d'action, de cause et d'effet **[Mizoguchi, 1997].**

III.5.1.3. Ontologies de domaine : Elles sont construites sur un domaine particulier de la connaissance. Les ontologies de domaine fournissent des vocabulaires au sujet des concepts dans un domaine et leurs relations au sujet des activités qui ont lieu dans ce domaine, et au sujet des théories et des principes élémentaires régissant ce domaine. Plusieurs ontologies de domaines existent déjà, telle que MENELAS15 dans le domaine médical. Entreprise16 est un autre exemple décrivant le domaine de l'entreprise **[Van Heijst& Schreiber &Wielinga, 1997].**

III.5.1.4. Ontologies de tâche : L'ontologie de tâche décrit les connaissances portant sur taches et/ou des activités particulières. Ces ontologies fournissent un ensemble de termes au moyen desquels on peut décrire au niveau générique comment résoudre un type de problème. Elles incluent des noms génériques (objectif, contrainte…), des verbes génériques (classer, sélectionner,…), des adjectifs génériques (assigné,..) et autres dans les descriptions de tâches **[N. Guarino, 1998]**.

III.5.1.5. Ontologies d'application : Aussi appelée ontologie de domaine-tache : Ce sont les ontologies les plus spécifiques, elles contiennent les connaissances requises pour une application particulière permettant ainsi de modéliser une activité spécifique dans un domaine donné **[Van Heijst& Schreiber &Wielinga, 1997].**

III.5.2. Les composants d'une ontologie :

III.5.2.1. Concept : ou classe, définissant un ensemble d'objet, abstrait ou concret, que l'on souhaite modéliser pour un domaine donné. Les connaissances portent sur des objets auxquels on se réfère à travers des concepts. Un concept peut représenter un objet matériel, une notion, une idée **[Uschold& King, 1995]**.

III.5.2.2. Les instances : ou individus, constituent la définition extensionnelle de l'ontologie (pour représenter les éléments spécifiques)

III.5.2.3. Les relations : Une relation permet de lier des instances de concepts ou des concepts génériques. Elles sont caractérisées par un terme ou plusieurs, et une signature qui précise le nombre d'instances de concepts que la relation lie, leurs types et l'ordre des concepts, c'est – à – dire la façon dont la relation doit être lue.

III.5.2.4. Les axiomes : Une ontologie est en outre composée d'axiomes qui forment des contraintes sémantiques pour le raisonnement et donnent un acompte d'une conceptualisation. Ils prennent la forme d'une théorie logique.

III.5.3. Thésaurus WordNet :

WordNet est une base de données lexicales construite par un groupe psychologues et de linguistes du laboratoire de sciences cognitives de l'université de Princeton, dirigé par le professeur Georges A. Miller. Elle a été initialement conçue dans le cadre d'un projet lancé en 1985 et gracieusement financé par l'agence de renseignements américaine (CIA), avec l'objectif de tester les déficits lexicaux dans des expériences de psychologie cognitive. A l'origine, ces concepteurs ne prétendaient construire ni une structure conceptuelle, ni une ontologie, mais bien une ressource lexicale rendant compte de l'usage des mots et de leur mise en relation dans la langue. Ce n'est qu'ensuite que le réseau lexical de WordNet a été perçu comme une représentation conceptuelle (Lexical Conceptual Graph ou LGC) [Guarino et al., 1999] qui pourrait tenir lieud'ontologie.

Comme impact, WordNet a inspiré d'autres projets tel EuroWordNet (lancé en 1996), qui construit une représentation conceptuelle indépendante des langues (interlangue), qui sert de noyau pour la majorité des langues européennes. Plusieurs autres propositions et initiatives ont aussi vu le jour ayant pour fin le "nettoyage" de parties de WordNet et leur attachement à des ontologies de haut niveau pour en faire une ontologie. Parmi ces travaux, on peut citer ceux de Guarino dans le projet ONTOCLEAN [Guarino et al., 2000] [Guarino et al.,2002] pour corriger les inadéquations dans les liens taxonomiques de WordNet,le projet ONION [Gangemi et al., 1999] où une structure sophistiquée est proposée pour regrouper plusieurs ontologies (ontologymerging) ou encore les travaux de Niles et Pease [Niles et al., 2003] pour relier (manuellement)une ontologie formelle de haut niveau appelée SUMO (SuggestedUpperMergedOntology) à WordNet.

Concernant le contenu de WordNet, il couvre la majorité des noms, verbes, adjectifs et adverbes de la langue Anglaise. Sa dimension ainsi que le domaine de la langue générale qu'il traite lui permettent souvent de couvrir les sujets traités dans les collections de

test conventionnelles de la RI (TREC, CLEF). Ces dernières sont le plus souvent de type presse.

WordNet a un réseau de 144 684 termes organisés en 109 377 nœuds (concepts) appelés Synsets (un exemple de noeud dans WordNet est donné). Le Tableau 1-1 donne des statistiques sur le nombre de mots et de concepts dans WordNet.

Catégorie	Mots	Concepts	Total Paires Mot-Sens
Nom	107 930	74 488	132407
Verbe	10 806	12 754	23255
Adjectif	21 365	18 523	31077
Adverbe	4 583	3 612	5721
Total	144 684	109 377	192460

Tableau 1 : Le nombre de mots et de concepts dans WordNet

Dans WordNet, une entrée est donc un concept qui est représenté par un Synset, c'est-à-dire l'ensemble des termes (mots ou groupes de mots) synonymes qui peuvent désigner ce concept. Les concepts reliés sémantiquement par une relation donnée à un Synset, sont représentés par une classe qui porte le nom de la relation. La relation de base entre les termes dans WordNet est la Synonymie. Les Synsets sont liés par des relations telles que spécifique-générique ou hyponyme-hyperonyme (is-a), et la relation de composition meronymie-holonymie (partie-tout) comme représentées dans le schéma de la Figure 4 .

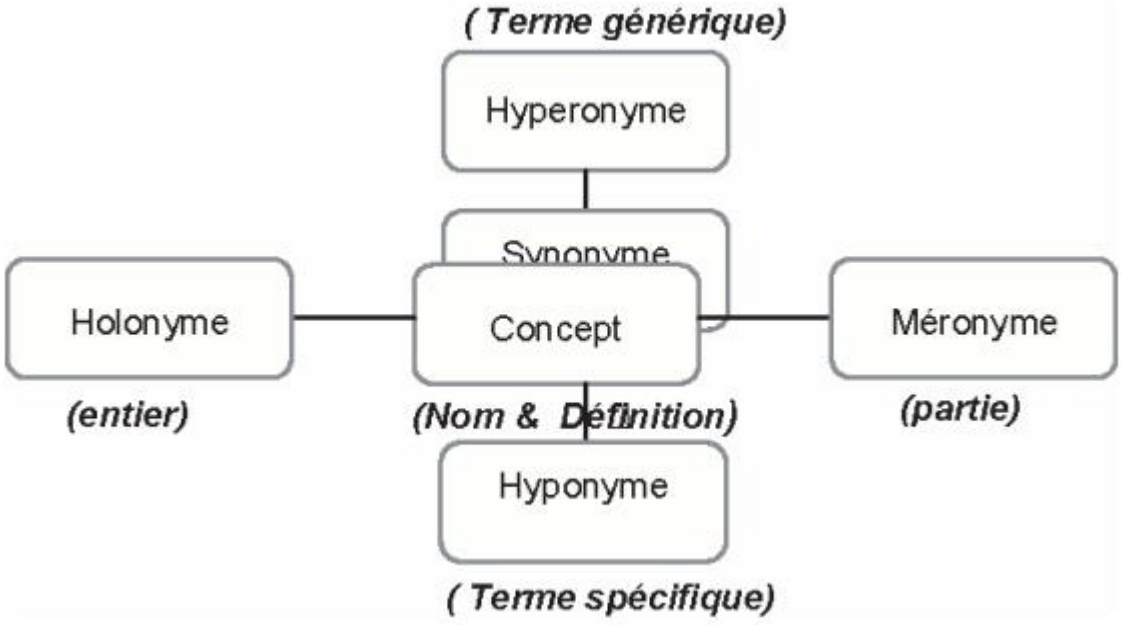


Figure 4 : Principales relations sémantiques dans WordNet

Ces relations peuvent être définies comme suit :

Synonymie, relie les termes représentant un même concept.

III.5.3.1. Relation Hyperonymie : C'est le terme générique utilisé pour désigner une classe englobant des instances de classes plus spécifiques. Y est un hyperonyme de X si X est un type de (kind of) Y.

III.5.3.2. Relation Hyponymie : C'est le terme spécifique utilisé pour désigner un membre d'une classe (relation inverse de Hyperonymie). X est un hyponyme de Y si X est un type de (kind of) Y.

III.5.3.3. Relation Holonymie : Le nom de la classe globale dont les noms meronymes font partie. Y est un holonyme de X si X est une partie de (is a part of) Y.

III.5.3.4. Relation Méronymie : Le nom d'une partie constituante (part of), substance de (substance of) ou membre (member of) d'une autre classe (relation inverse de l'holonymie). X est un méronyme de Y si X est une partie de Y. exemple : {voiture} a pour méronymes {{porte}, {moteur}}.

La Figure 5 donne un exemple de sous-hiérarchie extraite de WordNetcorrespondant au concept "car".

En plus de ces relations, on peut citer quelques autres relations qui sont cependant moins utilisées en pratique. C'est le cas notamment de la relation domain (rajoutée récemment pour la version WordNet 2.0), de la relation antonymy pour exprimer les sens opposés pour les synsets, de la relation entailment pour les verbes : Un verbe X entails (nécessite) Y si X ne peut être fait à moins que Y ne le soit

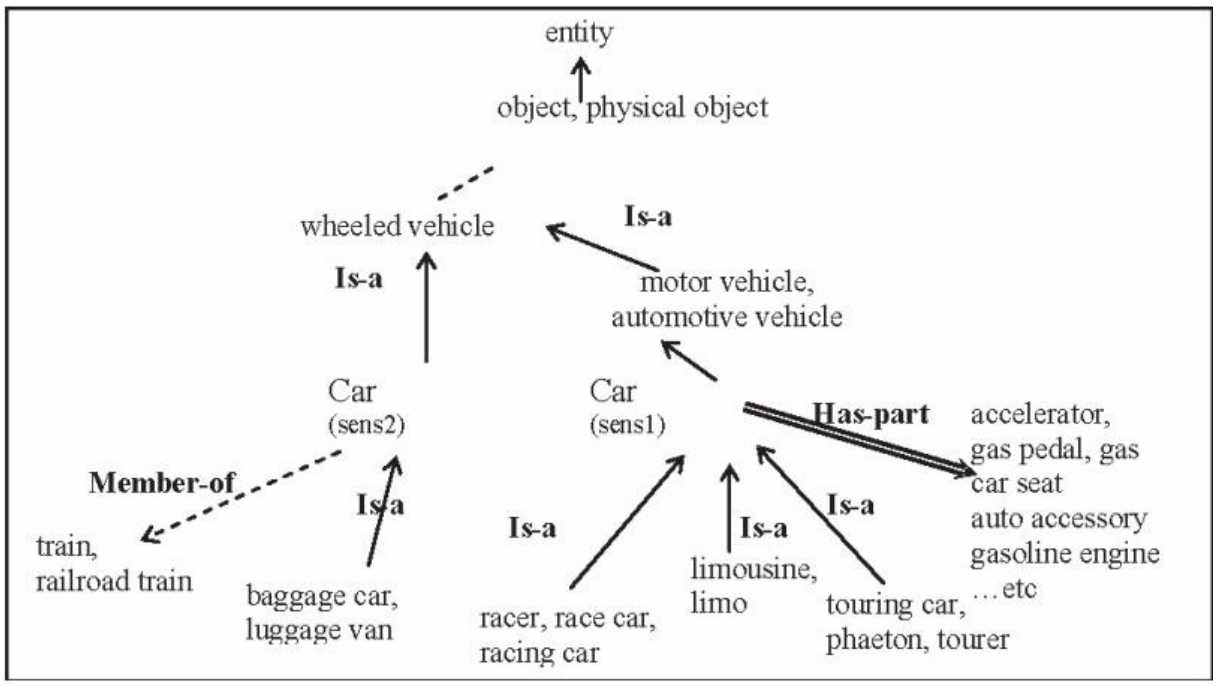


Figure 5 : Exemple de sous hiérarchie dans WordNet correspondent au concept "car"[Baziz, 2005]

IV. Les approches de désambiguïsation :

La désambiguïsation de sens de mot d'un texte ou Word SenseDisambiguation (WSD) consiste à déterminer le sens correct des mots de ce texte. Les approches de l'indexation sémantique s'appuient sur des algorithmes de désambiguïsation de mots (WSD), Nous présentons dans ce qui suit quelques approches proposées pour l'indexation sémantique.

IV.1. APPROCHE de Voorhees :

Voorhees [Voorhees, 1993] a créé un outil de désambiguïsation basé sur WordNet2. Pour désambiguïser une occurrence d'un mot ambigu, les synsets (sens) de ce mot sont classés en se basant sur la valeur de cooccurrence calculée entre le contexte de ce mot et un voisinage contenant les mots du synset dans la hiérarchie de WordNet. Voorhees a expérimenté cette approche sur une collection de test désambiguïsée (les requêtes de la collection de test sont aussi désambiguïsées manuellement) par rapport aux performances du même processus sur la même collection dans son état d'origine (ambigu). Les résultats de ses expérimentations ont montré que les performances du système de RI diminuent sensiblement dans le cas de l'utilisation des collections désambiguïsées *[ABBAS, 2014]*.

IV.2. APPROCHE de Mihalcea et al :

Mihalcea et Moldovan [Mihalcea and Moldovan, 2000] utilisent une méthode de désambiguïsation basée sur un corpus pré étiqueté sémantiquement (désambiguïsé) et WordNet. Un nouveau mot est désambiguïsé en tenant compte de sa relation avec les mots du corpus qui sont déjà désambiguïsés. Ce processus itératif leur permet d'identifier dans le corpus d'origine les mots qui peuvent être désambiguïsés avec une grande précision. Ils arrivent ainsi à désambiguïser 55% des mots (noms et verbes) avec une précision de 92%*[ABBAS, 2014]*.

IV.3. APPROCHE de Baziz :

Dans [BAZIZ, 2005], il a été proposé de reformuler une requête d'un utilisateur en s'appuyant sur une ontologie. L'approche consiste à projeter une requête sur une ontologie, identifier les nœuds (concepts) de l'ontologie qui représentent au mieux le contenu de la requête en utilisant différentes relations sémantiques, puis de la réécrire dans le format de départ (mots simples) avant de l'envoyer au SRI. Les résultats des expérimentations montrent que : - Les poids à affecter aux mots des concepts ajoutés à la requête suite à l'expansion, doivent être inférieurs à ceux des mots de la requête initiale (poids optimal =0.5). - Le nombre de termes issus d'un concept à retenir dans le processus d'expansion, doit être limité pour ne pas engendrer un bruit trop important. - La relation hyperonymie (généralisation) permet d'améliorer la précision globale (moyenne), tandis que la synonymie améliore la précision pour les premiers "meilleurs" documents retournés. Il est à noter que des schémas de pondération de concepts ont été proposés [Boubkeur, and Azzoug, 2013] afin de mieux exprimer l'apport sémantique d'un concept au contenu du document *[ABBAS, 2014]*.

IV.4. APPROCHE DE SCHÜTZ & PEDERSEN :

Schütz & Pedersen ont proposé une désambiguïsation basée seulement sur le corpus. Ainsi pour chaque mot à désambiguïser, on examine le contexte de chaque occurrence de ce mot dans le corpus. Les contextes similaires sont regroupés. Chacun de ces contextes similaires représente un sens individuel pour ce mot que Schütz & Pedersen désignent par usage de mot (Word uses) Affin de n'identifier que les sens fréquents d'un mot, un contexte similaire n'est identifié comme étant un sens que s'il apparaît plus d'une cinquantaine de fois dans le corpus, éliminant ainsi les sens les moins fréquents.

Pour désambiguïser une occurrence d'un mot, la technique consiste à classer les usages possibles d'un mot selon un score basé sur le recouvrement entre son contexte et les contextes d'usage. Une fois les mots désambiguïsés, la construction de leur index s'est faite de trois manières différentes. Dans le premier cas, une occurrence d'un mot est représentée simplement par le mot (le cas basique). Dans le second cas, par l'usage de mot le mieux classé. Et enfin, par une combinaison du mot et des n premiers usages de mots les mieux classés.

Leurs tests se sont effectués sur une collection relativement petite, qui est la collection TREC-1 catégorie B. Ils n'ont utilisé à cet effet que 25 requêtes en raison de la complexité du calcul

lors du regroupement des contextes similaires. Les résultats obtenus sont de l'ordre de 14 % de gain dans la précision. Le meilleur résultat correspond à la dernière représentation quand le nombre des premiers usages de mots utilisés est trois (3)*[ABBAS, 2014]*.

IV.5. APPROCHE DE KATZ ET AL. :

Dans une approche similaire à celle de Voorhees, Katz et al. [Katz &al., 1998] analysent les textes à indexer mot par mot. Chaque mot non vide rencontré est projeté sur WordNet dans l'objectif d'identifier le (ou les) synset(s) correspondant(s). Si un mot apparie plusieurssynsets, il est ambigu.

Pour désambiguïser, Katz et al proposent aussi une approche basée sur le contexte local. Le contexte local d'un mot est défini comme étant la liste ordonnée des mots démarrant du mot utile le plus proche du voisinage gauche ou droit jusqu'au mot cible. L'hypothèse de Katz et al., est que des mots utilisés dans le même contexte local (appelés sélecteurs), ont souvent des sens proches. Les sélecteurs des mots d'entrée sont extraits des contextes locaux gauche et droit, puis l'ensemble S de tous les sélecteurs obtenus est comparé avec les synsets de WordNet. Le synset qui a le plus de mots en commun avec S est sélectionné comme sens adéquat du mot cible*[ABBAS, 2014]*.

IV.6. APPROCHE de Lesk :

Présentée dans [Lesk 1986], est une méthode de désambiguïsation automatique supervisée dont le but est de discriminer les sens des mots polysémiques à l'aide d'un dictionnaire électronique, en l'occurrence, le *Oxford Advanced Learner'sDictionary of CurrentEnglish*. Le principe de base de cette méthode est de mesurer le chevauchement entre les différentes définitions, dans le dictionnaire, d'un mot ambigu et les définitions de ses voisins immédiats, dans une fenêtre de 10 mots.

Conclusion :

Dans ce chapitre, nous avons posé la problématique générale de l'indexation classique. Et pour solution on a introduit l'indexation sémantique comme une approche basée sur les sens des mots, qui sont extraits des ressources sémantiques telles que les ontologies, thésaurus et taxonomies.

Nous avons aussi présenté quelques différentes approches de désambiguïsation, qui utilise ces ressources, qui permettent d'identifier les concepts à partir d'un document textuel, et qu'ils offrent aussi des connaissances sur la corrélation entre les concepts à l'aide des mesures de similarité que nous allons présenter dans le prochain chapitre.

Chapitre III : Calcul de similarité

Introduction :

Lorsque l'on entend parler de calcul de similarité en informatique, c'est souvent dans le cadre de données textuelles : comment classer, regrouper des documents ? Comment, pour une requête, retourner la liste des documents les plus pertinents (dans le cas d'un moteur de recherche par exemple) ?

La notion de similarité sémantique (distance sémantique) est utilisée pour exprimer la ressemblance entre des concepts. En recherche d'information conceptuelle, les mesures de similarité jouent un rôle important, en particulier dans le processus de désambiguïsation des termes, la pondération des concepts et l'évaluation de la pertinence. L'objectif des mesures de similarité sémantique est d'estimer la ressemblance entre les documents.

L'objectif de ce chapitre est de présenter les bases de la similarité. Nous présentons certaines approches permettant de comparer des textes. Les approches présentées ont été sélectionnées pour répondre au mieux au contexte (détaillé dans la section suivante). Ainsi, ce chapitre ne prétend pas donner une liste exhaustive de toutes les méthodes existantes, mais tente de donner un aperçu des méthodes les plus utilisées dans le contexte de notre étude.

I. Similarité entre textes :

Évaluer la similarité entre documents textuels est une des problématiques importantes de plusieurs domaines comme la recherche d'information ou l'extraction de connaissances à partir de données textuelles (Text Mining), dans le domaine de détection plagiat et dans le domaine de classification des documents .Dans chacun de ces domaines, les similarités sont utilisées pour différents traitements :

- En analyse de données textuelles, les similarités sont utilisées pour la description et l'exploration de données ;

- En recherche d'information, l'évaluation des similarités entre documents et requêtes est utilisée pour identifier les documents pertinents par rapport à des besoins d'information exprimés par les utilisateurs ;

- en Text Mining, les similarités sont utilisées pour produire des représentations synthétiques de vastes collections de documents.

II. La recherche d'information et la similarité (document/requête) :

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 [Daoud., 2009].
Elle a pour objet de sélectionner dans une collection les informations pertinentes répondant à des besoins d'utilisateurs.

II.1. Les concepts de base de la RI :

II.1.1. Systèmes de recherche d'information (SRI) :

Un système de recherche d'information (SRI) est un système qui a pour rôle de d'automatiser la tache de la RI mettre en œuvre des techniques et des moyens permettant de retourner les documents pertinents d'une collection en réponse à un besoin en information d'un utilisateur, exprimée par un langage de requêtes qui peut être le langage naturel, une liste de mots clés ou un langage booléen. [Baeza., 2011]
Cette définition met en évidence plusieurs concepts clés que nous allons expliciter dans ce qui suit: Document, Collection de documents, Requête, Pertinence et Besoin d'information.

1. Document :

Le document représente le conteneur élémentaire d'information, exploitable et accessible par le SRI. Un document peut être un texte, une page WEB, une image, une bande vidéo, etc. Dans notre contexte, nous appelons document toute unité qui peut constituer une réponse à un besoin en information exprimé par un utilisateur.

2. Collection de document :

La collection de documents (ou fond documentaire, corpus) constitue l'ensemble des informations exploitables et accessibles. Elle est constituée d'un ensemble de documents. Dans le cas général et pour un souci d'optimalité, la base constitue des représentations simplifiées mais suffisantes pour ces documents. Ces représentations sont étudiées de telle sorte que la gestion (ajout suppression d'un document) ou l'interrogation (recherche) de la base se font dans les meilleurs conditions de coût.[Baziz., 2005]

3. Requête :

La requête est une expression approximative du besoin en information de l'utilisateur, Elle représente l'interface entre le SRI et l'utilisateur, et il existe plusieurs langage pour formuler une requête, nous citrons :

1. langage booléen : L'utilisateur exprime sa requête sous forme d'un ensemble de termes reliés entre eux par des opérateurs booléens (ET, OU, NON), cas du système DIALOG.

2. langage naturel : La requête de l'utilisateur est exprimée en langage libre (Naturel) sous forme de mots clés, le SRI se charge d'analyser, de reformuler et de traduire ces mots clés en une requête utilisable et compréhensible par ce système, cas des systèmes SMART.

3. langage graphique : Une interface d'aide à la formulation de la requête est proposée à l'utilisateur. En effet, une vue d'ensemble de la base d'information et en particulier une vue de termes représentant le contenu sémantique des documents, est donnée a l'utilisateur pour l'assister à formuler sa requête, cas du système NEURODOC.

4. Pertinence :

La pertinence est une notion fondamentale en RI et fait l'objet de SRI, elle dénote une relation reliant une requête utilisateur à un document qui satisfait le besoin en information visé par cette dernière, plusieurs définitions [Saracevic., 1970] lui sont rattachées, elle peut être définie comme la correspondance entre un document et une requête ou encore comme une mesure d'informativité du document à la requête.

On peut distinguer deux types de pertinence : *la pertinence système* et *la pertinence utilisateur.*

***1. La pertinence Système* :** est souvent présentée par un score attribué par le SRI afin dévaluer l'adéquation du contenu des documents vis-à-vis de celui de la requête. Ce type de pertinence est objectif et déterministe. [Cleverdom., 1970]

***2. Pertinence utilisateur* :** se traduit par les jugements de pertinence de l'utilisateur sur les documents fournis par le SRI en réponse à une requête. La pertinence utilisateur est subjective, car pour un même document retourné en réponse à une même requête, il peut être jugé différemment par deux utilisateurs distincts (qui ont des centres d'intérêt différents). De plus, cette pertinence est évolutive, un document jugé non pertinent à l'instant t pour une requête peut être jugé pertinent à l'instant t+1, car la connaissance de l'utilisateur sur le sujet a évolué. [Harter., 1992] [Mezzaro., 1997] [Saracevic., 1996]

Le but de tout système de recherche d'information est alors de faire correspondre au mieux la pertinence système avec la pertinence utilisateur [Denos., 1997].

5. Besoin d'information :

La notion de besoin en information en recherche d'informations est souvent assimilée au besoin de l'utilisateur. Trois types de besoin utilisateur ont été définis par [Ingwersen., 1994] :

1. Besoin vérificatif : l'utilisateur cherche à vérifier le texte avec les données connues qu'il possède déjà. Il recherche donc une donnée particulière, et sait même souvent comment y accéder. La recherche d'un article sur Internet à partir d'une adresse connue serait un exemple d'un tel besoin. Un autre exemple serait de chercher la date de publication d'un ouvrage dont la référence est connue. Un besoin de type vérificatif est dit stable, c'est-à-dire qu'il ne change pas au cours de la recherche.

2. Besoin thématique connu *:* l'utilisateur cherche à clarifier, à revoir ou à trouver de nouvelles informations dans un sujet et domaine connus. Un besoin de ce type peut être stable ou variable ; il est très possible en effet que le besoin de l'utilisateur s'affine au cours de la recherche. Le besoin peut aussi s'exprimer de façon incomplète, c'est-à-dire que l'utilisateur n'énonce pas nécessairement tout ce qu'il sait dans sa requête mais seulement un sous-ensemble. C'est ce qu'on appelle dans la littérature le label.

3. Besoin thématique inconnu *:* cette fois, l'utilisateur cherche de nouveaux concepts ou de nouvelles relations hors des sujets ou domaines qui lui sont familiers. Le besoin est intrinsèquement variable et est toujours exprimé de façon incomplète.

Le processus de la RI est constitué de plusieurs étapes qui sont : l'indexation, le choix de modèle de recherche d'information, la reformulation de la requête et de l'Appariement Document-Requête.

L'une des étapes clés dans le processus de RI est l'appariement document-requête. Il consiste à mettre en correspondance la représentation de la requête avec les représentations des documents. Cette correspondance est mesurée par un score de pertinence associé au document, qui reflète son degré de ressemblance ou de similarité avec la requête. Dans le modèle de recherche vectoriel (le plus simple et le plus intuitif des modèles de RI), documents et requêtes sont représentés par des vecteurs de termes pondérés. L'appariement est alors basé sur une mesure de similarité entre les vecteurs correspondants, généralement calculée comme le cosinus [Salton et al., 83] entre les deux vecteurs. Cette mesure dépend des termes communs au document et à la requête.

Les documents sont représentés à l'aide d'annotations, c'est-à-dire des graphes de concepts issus d'une ontologie. Une annotation peut être constituée d'un ou plusieurs graphes de concepts.

Le calcul de similarité entre annotations intervient donc à trois niveaux :

- au niveau des annotations ;
- au niveau des graphes de concepts ;
- et au niveau des concepts.

Il est donc nécessaire de disposer d'une mesure de similarité pour chacun de ces niveaux. La similarité entre annotations utilise la similarité entre graphes de concepts, cette dernière reposant sur la similarité entre concepts. Nous détaillons ces mesures dans les sections suivantes.

II.1. les méthodes ProxiGénéa :

Notre mesure s'inspire du principe d'arbre généalogique familial. Nous considérons en effet que les termes d'un vocabulaire peuvent être organisés sous la forme d'un arbre hiérarchique où les termes les plus spécifiques sont rattachés aux termes plus génériques par une relation père-fils.

La similarité entre deux annotations est déterminée par la moyenne des similarités des graphes de concepts qui les constituent. L'algorithme de calcul est le suivant :

```
simAnnotation ← 0 ;
Pour chaque graphe de concepts G1 de l'annotation A1 faire
        simGrapheMax ← 0 ;
        Calculer la similarité simGraphes G1 et G2
  Si simGraphes(G1,G2) > simGrapheMax Alors
            simGrapheMax ← simGraphe(G1,G2);
  Fin Si ;
        simAnnotation ← simAnnotation + simGrapheMax ;
```

Fin Pour ;
simAnnotation ← simAnnotation / NombreGraphes(A1) ;

Nous décrivons dans la section qui suit le principe de calcul de la similarité entre deux graphes de concepts, utilisé pour déterminer la similarité entre annotations.

II.1.1. Similarité entre graphes de concepts :

La similarité entre deux graphes est définie comme la moyenne pondérée des similarités entre les concepts qui les composent.

Deux concepts sont comparables s'ils sont descendants d'un même top-concept. Les top-concepts sont les concepts fils de la racine de l'arbre taxonomique. Il s'agit donc des concepts les plus génériques de l'ontologie.

Par ailleurs, les concepts peuvent avoir des importances différentes en fonction des applications. Ce degré d'importance est défini pour chaque top-concept de l'ontologie arbitrairement ou après une phase d'apprentissage. Le degré d'importance d'un concept correspond à celui du top-concept dont il est le descendant.

Soient :

- G1 et G2 deux graphes de concepts ;
- Nœuds(G) l'ensemble des nœuds (i.e. les concepts) du graphe G ;
- G1 i et G2 j des concepts appartenant respectivement aux graphes G1 et G2 ;
- Coef(G i) la fonction déterminant le degré d'importance d'un concept du graphe G ;
- et SimConcept(G1 i ,G2 j) la similarité entre les concepts G1 i et G2 j .

La similarité entre deux graphes de concepts est définie ainsi :

$$SimGraphes(G1,G2) = \frac{\sum_{i=1}^{|Noeuds(G1)|} Coef(G1_i) \cdot \max_{j=1}^{|Noeuds(G2)|} (SimConcept(G1_i, G2_j))}{\sum_{i=1}^{|Noeuds(G1)|} Coef(G1_i)}$$

SimConcept peut être déterminée par plusieurs mesures de similarité, comme celle proposée par Wu et Palmer [11]. Dans la section suivante, nous proposons

ProxiGénéa, une mesure de similarité alternative.

Les documents et la requête sont d'abord associés à des modèles de représentation qui vont servir de base au calcule de similarité.

II.2. Le modèle LSI/PLSI :

Le modèle Latent Semantic Indexing (LSI) découle du modèle vectoriel standard. Il tente de prendre en considération la structure sémantique des termes pour la représentation des documents. Dans ce modèle, les documents sont représentés dans un espace réduit de termes d'indexation [Dee&al 1990], [Sch&al 1995],[Lan&al1998]. Les techniques LSI utilisent, dans un premier temps, une matrice M (documents × unités linguistiques), dans laquelle chaque élément Wij est une pondération en fonction du nombre d'occurrences du terme Tj dans le document Di. Soit n le nombre de documents de la collection et t le nombre des termes d'indexation. La matrice M peut être représentée comme suit.

$$M = \begin{pmatrix} d_1 \\ d_2 \\ \dots \\ d_n \end{pmatrix} = \begin{pmatrix} W_{11} & W_{12} & \dots & W_{1t} \\ W_{21} & W_{22} & \dots & W_{2t} \\ \dots & \dots & W_{ij} & \dots \\ W_{n1} & W_{n2} & \dots & W_{nt} \end{pmatrix}$$

Une décomposition en valeurs singulières (SVD) de la matrice M est ensuite effectuée. Après cette décomposition, seuls les k premiers vecteurs propres sont intégrés, les axes factoriels correspondant aux plus grandes valeurs propres sont conservés en application du théorème de Eckart et Young [Dee&al 1990]. Ce théorème montre qu'il s'agit dans ce cas des axes permettant un ajustement dans un espace de dimension réduite minimisant la perte d'information. Dans LSI, la valeur représentée dans la matrice sur laquelle est appliquée la décomposition est définie comme étant le produit du poids local du terme, i.e. poids du terme dans le document, et du poids global du terme, i.e. poids du terme dans la collection de documents. Cette valeur peut aussi correspondre à la fréquence d'un terme donné dans un document donné. Ces valeurs ne sont que des pondérations proches de TF/IDF [Dee&al 1990], [Lan&al 1998].

Hoffmann a proposé un modèle probabiliste du Latent Semantic Indexing (PLSI). Il considère l'hypothèse que les documents sont associés à un certain nombre de sens (Latents) et que les termes correspondent à l'expression de ces sens [Hof 1999]. De façon probabiliste, notant W l'ensemble des mots, D l'ensemble des documents, tel que : W = {W1,…,Wt} et

D = {D1,…,Dn}, la probabilité de la paire observée ($d \in D$, $w \in W$) est donnée par la formule suivante :

$$p(d, w) = p(w \mid d)\, p(d)$$

Cette probabilité est calculée en utilisant un algorithme de type Expectation-Maximization (EM) [Dem&al 1977]. Les dimensions de l'espace réduit du modèle LSI correspondent ici aux sens du modèle PLSI.

II.3. Le modèle DSIR :

Le modèle de représentation de textes Distributional Semantic for Information Retrieval (DSIR) [Raj&al 2000], [Bes&al 2001] est basé sur les cooccurrences des mots dans les collections de documents. Les contextes des unités linguistiques sont des éléments essentiels du modèle DSIR car ils constituent le support principal pour la dérivation des représentations des mots. Ces dernières sont obtenues à partir des fréquences de cooccurrences entre l'ensemble des mots d'une base documentaire et les termes d'indexation de cette dernière. Dans ce contexte, un mot (unité linguistique) Ui est représenté par un vecteur de poids associés aux fréquences de cooccurrences de ce mot avec l'ensemble des termes d'indexation T. Ce vecteur est appelé profil de cooccurrences de Ui par rapport à T. Notons que dans le modèle DSIR un contexte de cooccurrence correspond à une phrase.

Soit Wij le poids associé à la fréquence de cooccurrence de Ui avec un terme d'indexation Tj et t = |T|. Ui est représenté par un vecteur VCOi :

$$VCOi = (Wi1,\ldots,Wij,\ \ldots,Wit)$$

Ainsi, une collection de documents est représentée par une matrice M de cooccurrences (unités linguistiques × termes d'indexation).

Soit U l'ensemble des mots de la collection et n le nombre de ces mots. Soit T l'ensemble des termes d'indexation et t le nombre de ces derniers. La matrice M est définie comme suit.

$$M = \begin{pmatrix} U_1 \\ U_2 \\ \ldots \\ U_n \end{pmatrix} = \begin{pmatrix} W_{11} & W_{12} & \ldots & W_{1t} \\ W_{21} & W_{22} & \ldots & W_{2t} \\ \ldots & \ldots & W_{ij} & \ldots \\ W_{n1} & W_{n2} & \ldots & W_{nt} \end{pmatrix}$$

Le modèle DSIR utilise une fonction de représentation d'un document à partir de la matrice de distribution des unités linguistiques. Dans ce cas, il s'agit de la matrice M. Dans un premier temps, un document est représenté par un vecteur V d'occurrence suivant le modèle vectoriel standard. Puis ce vecteur est transformé via la matrice M par une multiplication du vecteur V par M. Cette représentation prend en considération la distribution des cocitations de mots et intègre des connaissances syntaxiques dans la phase de sélection des cooccurrences. En effet, les cooccurrences des mots sont prises entre les gouverneurs des groupes syntaxiques ou les constituants d'un même groupe syntaxique. Une étape de filtrage syntaxique est donc nécessaire avant toute représentation de textes.

III. Le domaine de détection plagiat :

Avec l'essor du Web, la production et la diffusion de documents se sont développées de manière exponentielle. La pratique du Copier/Coller (plagiat) se standardise sans pour autant faire référence aux auteurs.

Le «plagiat». A l'origine, du latin plagiarius, qui signifiait dans la Rome antique le fait de voler l'esclave d'un autre ou de vendre une personne libre.

«Selon la section Droit d'auteur et plagiat du site Infosphère de l'Université de Montréal, plagier, c'est : s'approprier le travail créatif de quelqu'un d'autre et le présenter comme sien; s'accaparer des extraits de texte, des images, des données, etc. provenant de sources externes et de les intégrer à son propre travail sans en mentionner la provenance;

résumer l'idée originale d'un auteur en l'exprimant dans ses propres mots, mais en omettant d'en mentionner la source.» On voit que la détection automatique du plagiat est une tâche ardue qu'aucun système informatique actuel ne peut mener à bien en son entier. Il faudrait pour cela être capable de reconnaître qu'un texte emprunte des idées à un autre, ce qu'on est assez loin de savoir faire de manière fiable. C'est pourquoi les logiciels dits de «détection de plagiat» se limitent essentiellement à mettre en évidence des indices de plagiat, par exemple l'apparition de la même séquence de mots dans deux textes.

III.1. Détection de plagiat :

L'idée de créer des logiciels de détection de plagiat n'est pas récente. En effet, dans les années 1970-1980 des professeurs d'informatique avaient déjà développé des logiciels sophistiqués pour tenter de détecter le plagiat dans les travaux de programmation que devaient rendre les étudiants. Les premiers documents électroniques considérés étaient donc des programmes informatiques écrits dans une langue artificielle (FORTRAN, Pascal, etc.). La généralisation de l'usage du traitement de texte, dans les années 1980-1990 suscita la création de logiciels de comparaison et de détection prenant en compte les langues naturelles. On réutilisa pour cela les techniques et modèles développés dès les années 1960 par G. Salton pour la recherche d'informations et on créa également de nouvelles techniques et mesures de similarité et de classement de documents. **[Michelle Bergadaà. ;2008]**

Avec l'arrivée de l'Internet puis du World Wide Web le problème prend une tout autre dimension car on doit passer de la détection locale (copie entre étudiants, récupération de travaux des années précédentes, etc.) à une détection globale sur le Web et ses milliards de documents. Il a fallu adapter les techniques de détection à cette problématique de masse.

La question s'est donc posée rapidement pour nombre de créateurs de savoir comment se servir de la mutation des pratiques qu'allait entraîner la révolution vécue à un niveau individuel et des profits économiques qui étaient envisageables. Ainsi a-t-on vu se développer, d'une part, des propositions visant à détecter la fraude de l'écrit et, d'autre part, des sites offrant des documents que l'on pouvait acheter et faire passer pour siens en toute qui étude.

III.2. Les principes de la détection de similarités et de plagiat :

On peut dire que les logiciels de «détection de plagiat» savent, actuellement, tester si le texte A et le texte B ont des chaînes de mots en commun. Il s'agit donc d'une opération purement syntaxique qui ignore le sens des textes.

De tels logiciels sont aujourd'hui indispensables, car on comprend aisément qu'il n'est absolument pas envisageable de comparer un document suspect successivement avec chacun des milliards de documents du Web. Il faut donc utiliser des techniques dites d'indexation qui permettent de trouver rapidement les documents qui contiennent une liste de mots cherchés. On constitue pour cela des index qui sont des bases de données associant à chaque mot de la langue la liste de tous les documents du Web où ce mot apparaît au moins une fois. Chercher

un ensemble de mots revient alors à consulter les listes de documents correspondant à ces mots et à en faire l'intersection.

Il existe déjà de bons index du Web et des moteurs de recherche associés (Google, Yahoo, Exalead…), et la méthode de détection la plus simple à mettre en œuvre consiste à utiliser l'un de ces moteurs en lui soumettant successivement chaque phrase ou partie de phrase du document à vérifier. En principe, on peut facilement créer un logiciel qui effectue automatiquement ces opérations et crée un rapport de similarité en collectant les réponses du ou des moteurs utilisés. En pratique, cette méthode ne fonctionne pas car elle conduirait à surcharger les moteurs de recherche (l'analyse d'un seul texte de quelques dizaines de pages générerait des centaines voire des milliers de requêtes). A l'heure actuelle, les moteurs de recherche se prémunissent contre ce type d'utilisation en interdisant qu'on leur soumette plus d'un nombre donné de requêtes par unité de temps et par utilisateur.

Le deuxième type de méthodes de détection consiste à utiliser un des outils développés spécifiquement à cet usage par des entreprises privées. Les outils les plus connus (Turnitin, Compilatio, etc.) ont créé leur propre index du Web ou d'une partie du Web. La constitution d'un tel index nécessite évidemment des moyens importants: il faut parcourir régulièrement des milliers de sites, récupérer leurs documents, les stocker dans une base de données et constituer un index par mots. C'est ainsi que l'expression «achat d'un logiciel anti-plagiat» est impropre. Lorsqu'une institution se procure l'accès à un tel logiciel de détection, elle n'achète pas un logiciel autonome, mais le droit d'utiliser un logiciel qui est autorisé à accéder à l'index constitué par la société éditrice du logiciel. **[Michelle Bergadaà. ;2008]**

III.3. La notion de similitudes :

On distingue plusieurs types de similitudes allant de la ressemblance jusqu'à l'identité même [Simac-Lejeune., 2013b]. Les ressemblances sont les types de similitudes les plus difficiles à repérer et sont pour cause le point faible des logiciels anti-plagiat actuels. Dans notre cas, on distingue trois types majeurs de similitudes textuelles, de la plus simple à détecter à la plus complexe:

- **la copie**, qui consiste à copier mot à mot tout ou partie d'un texte dans un autre. Pour exemple, considérons la phrase suivante présente dans un texte : *« En cinquante ans, grâce à des efforts considérables dans la recherche et l'élaboration de la fusion, la performance des plasmas a été multipliée par 10'000. »* Elle sera recopiée à l'identique dans un autre texte .
- **la paraphrase**, aussi appelée reformulation paraphrastique, qui consiste à reprendre une phrase d'un texte pour la détailler ou l'expliciter. Elle conserve donc l'ordre des éléments évoqués, autorisant simplement le changement de vocabulaire, l'ajout, la suppression et la substitution de mots. Toujours en considérant la phrase de l'exemple précédent, une paraphrase possible serait : *« En une cinquantaine d'années, grâce à un immense effort de recherche, la performance des plasmas produits par les machines de fusion a été multipliée par 10000. »* .On remarque la conservation des concepts, mais aussi la substitution ou la suppression de certains d'entre eux.

- **la reformulation**, qui autorise toutes modifications textuelles à condition que le sens de la phrase soit conservé. Cela donne souvent lieu à un changement d'ordre des concepts. La

reformulation de la phrase exemple serait : *« La performance des plasmas produits par les machines de fusion a été multipliée par 10,000 grâce à un immense effort de la recherche bien que cela ait pris une cinquantaine d'années. ».*

III.4. Les approches de détection de plagiat :

Lorsque les processus anti-plagiat comparent deux documents, ils recherchent les éléments de l'un également présents dans l'autre. Ils tentent de détecter des similitudes, toutes informations communes laissant penser qu'un plagiat a pu avoir lieu. La comparaison mot à mot est certes efficace pour trouver les zones de « copier/coller » mais les plagiaires ne se contentent plus de copier des éléments depuis une source, ils essaient à présent de camoufler leurs emprunts d'idées derrière des modifications syntaxiques. Face à ces situations différentes approches sont apparues pour détecter le plagiat **[Jérémy & Alain., 2015] :**

- **les approches stylométriques :** [Iyer et Singh., 2005] qui suggèrent qu'en analysant des statistiques de fréquences de mots ou bien d'autres caractéristiques d'un texte on peut en reconnaître l'auteur, et ainsi, si un passage du document ne possède pas les mêmes caractéristiques que le reste du document, on peut en déduire que ce passage aura été emprunté à un autre auteur [Oberreuter.,2013] et [Velásquez.,2013], [van Halteren., 2004], [Jardinoet al., 2007].

- **les approches de calcul de distances :** [Simac-Lejeune, 2013a] qui propose de calculer une distance « sémantique » entre deux textes après avoir extrait les mots clefs de chaque texte, exposant ainsi l'emprunt probable de l'un dans l'autre.

- **Les approches par alignement :** Les approches les plus répandues sont les méthodes par alignement [Callison-Burch et al., 2008]; [Bannard., 2005 [et [Callison-Burch., 2005]. Servant la plupart du temps dans un contexte bi-linguale (alignement d'un texte et de sa traduction), elles consistent à aligner deux textes par leurs mots ou groupes de mots en communs et ainsi de repérer les mots ou groupes de mots différents mais équivalents. Certaines recherches [Shen et al., 2006], visant à produire des paraphrases, se sont également avérées intéressantes. En effet, étudiant la possibilité de générer automatiquement des paraphrases, un processus d'assemblage puis de désassemblage s'est dégagé, remettant ainsi sur le devant de la scène les méthodes d'alignement. Proche de ces méthodes avec alignement, on peut citer le travail de Fenoglio et al. (2007) traitant de la comparaison de versions de documents textuels à la façon des serveurs de versions. Il met en lumière les transformations élémentaires (déplacements, insertions, suppressions et remplacements de blocs de caractères), identifiées depuis longtemps par les spécialistes de la génétique textuelle [Biasi., 2000], [Grésillon., 1994] comme éléments fondateurs d'une paraphrase.

IV. Le domaine de classification des documents :

La classification des documents a pour objectif de regrouper les documents similaires, c'est à dire thématiquement proches, au sein d'un même ensemble. L'intérêt d'une telle démarche est d'organiser les connaissances de façon à pouvoir effectuer, par la suite, une recherche ou une extraction d'information efficace.

IV.1. Les approches de classification des documents ; La classification automatique consiste à regrouper divers objets (les individus) en sous-ensembles d'objets (les classes).

On distingue dans le domaine de la classification automatique deux types d'approches, la classification supervisée et la classification non supervisée.

1. La classification supervisée : les classes sont connues à priori, elles ont en général une sémantique associée. Elle suppose qu'il existe déjà une classification de documents. C'est le cas par exemple d'une bibliothèque ou d'un moteur de recherche comme Yahoo. Le but est alors de classer automatiquement un nouveau document.

Soit D= $\{d_1, d_2,\ldots,d_i,\ldots,d_m\}$ un ensemble de documents représentés chacun par une description :

$$\vec{d_1},\vec{d_2},\ldots,\vec{d_m}$$

Et $C=\{C_1,C_2,\ldots,C_k,\ldots,C_c\}$ un ensemble de classes, la classification supervisée suppose connues deux fonctions. La première fait correspondre à tout individu d_i une classe C_k. Elle est définit au moyen de couples (d_i,C_K) donnés comme exemples au système. La deuxième fait correspondre à tout individu d_i sa description.

La classification supervisée consiste alors à déterminer une procédure de classification :

$$C^f : \vec{d_i} \rightarrow C_k$$

Qui à partir de la description de l'élément détermine sa classe avec le plus faible taux d'erreurs. La performance de la classification dépend notamment de l'efficacité de la description.

1.1. Méthodes d'apprentissage supervisé : Comme les documents sont nombreux ou que leur nombre augmente sans cesse, il serait difficile de programmer à l'avance des règles de décision pour déterminer la classe d'un nouveau document. Même si cela était possible, ces règles devraient être régulièrement modifiées par l'utilisateur pour qu'elles reflètent la réalité actuelle. Nous présentons donc des méthodes d'apprentissage qui, à partir de documents déjà classés. Il existe de nombreuses méthodes d'apprentissage supervisé :

1.1.1. K plus proches voisins *(k-NN)* : L'idée de k-NN est de représenter chaque texte dans un espace vectoriel, dont chacun des axes représente un élément textuel. Cette méthode diffère des traditionnelles méthodes d'apprentissage car aucun modèle n'est induit à partir des exemples. Les données restent telles quelles : elles sont simplement stockées en mémoire. Pour prédire la classe d'un nouveau cas (où ranger un nouveau

document ?), l'algorithme cherche les K plus proches voisins de ce nouveau cas et prédit (s'il faut choisir) la réponse la plus fréquente de ces K plus proches voisins. La méthode utilise donc deux paramètres : le nombre K et la fonction de similarité pour comparer le nouveau cas aux cas déjà classés.

$$d(x_i, x_j) \equiv \sqrt{\sum_{r=1}^{n} (a_r(x_i) - a_r(x_j))^2}$$

Un inconvénient majeur de k-NN reste le temps qu'il met pour classer un nouveau cas : il faut calculer la similarité entre K (ou même les N) exemples et le nouveau cas, puis décider quelle classe choisir (soit par majorité, soit en fonction de pondération selon la distance de chaque exemple avec le nouveau cas).

1.1.2. Arbres de décisions :

Les arbres de décision représentent une méthode très efficace d'apprentissage supervisé. Il s'agit de partitionner un ensemble de données en des groupes les plus homogènes possible du point de vue de la variable à prédire. On prend en entrée un ensemble de données classées, et on fournit en sortie un arbre qui ressemble beaucoup à un diagramme d'orientation où chaque nœud final (feuille) représente une décision (une classe) et chaque nœud non final (interne) représente un test. Chaque feuille représente la décision d'appartenance à une classe des données vérifiant tous les tests du chemin menant de la racine à cette feuille.

L'exemple suivant montre un ensemble de données avec quatre attributs : Ensoleillement, Température, Humidité, Vent et l'attribut à prédire Jouer.

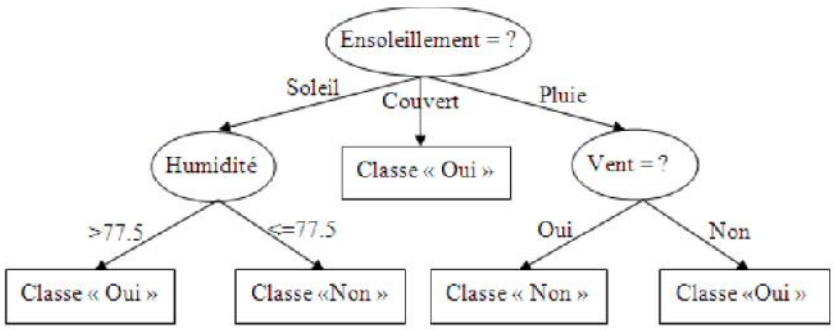


Figure 6 : exemple d'arbre de décision

En effet, toutes les données ayant l'attribut Ensoleillement="Soleil" et l'attribut Humidité>77.5 appartiennent à la classe 1 ("oui"). Toute nouvelle donnée peut être classée en testant ses valeurs d'attributs l'un après l'autre en commençant de la racine jusqu'à atteindre une feuille c'est-à-dire une décision. Pour construire un tel arbre, plusieurs algorithmes existent : ID3, CART, C4.5,...etc. On commence généralement par le choix d'un attribut puis le choix d'un nombre de critères pour son nœud. On crée pour chaque critère un nœud concernant les données vérifiant ce critère. L'algorithme continue d'une façon récursive jusqu'à obtenir des nœuds concernant les données de chaque même classe.

Les algorithmes de constructions d'arbre de décision les plus utilisés sont les Algorithme ID3.

1.1.3. Naïve Bayes (ou Simple Bayes) :

Nommés d'après le théorème de Bayes, ces méthodes sont qualifiées de "Naïve" ou "Simple" car elles supposent l'indépendance des variables. L'idée est d'utiliser des conditions de probabilité observées dans les données. On calcule la probabilité de chaque classe parmi les exemples. Ce sont les "prior probabilities". Par exemple, si la classe "informatique" revient 2 fois sur les 5 documents donnés en exemple, sa "prior probability" sera de 2/5. En plus des "prior probas", l'algorithme calcule les fréquences d'apparition de chaque variable d'entrée avec celles de sortie. Pour classer des documents, les variables d'entrée sont les mots présents dans l'ensemble des documents. A chaque mot on calcule le nombre de fois qu'il apparaît dans les documents classés dans une classe donnée. On calcule cette fréquence pour chaque classe.

1.1.4. Réseaux de neurones :

Les réseaux de neurones sont utilisés pour leur capacité à apprendre à partir d'exemples bruités comme les caméras ou les micros (reconnaissance de forme ou de son). Mais ils sont aussi utilisables pour des problèmes où les méthodes symboliques (arbres de décisions) sont souvent utilisées. Leur performance est alors équivalente.

Les réseaux de neurone sont appropriés lorsque le temps d'apprentissage n'est pas essentiel : ce temps est en effet souvent très supérieur à d'autres méthodes comme les arbres de décision. Par contre, la classification d'un nouveau cas (par exemple un document) est très rapide.

1.1.5. Machines à support de vecteurs (ou SVM) :

Cette technique - initiée par Vapnik - tente de séparer linéairement les exemples positifs des exemples négatifs dans l'ensemble des exemples. Chaque exemple doit être représenté par un vecteur de dimension n. La méthode cherche alors l'hyperplan qui sépare les exemples positifs des exemples négatifs, en garantissant que la marge entre le plus proche des positifs et des négatifs soit maximale. Intuitivement, cela garantit un bon niveau de généralisation car de nouveaux exemples pourront ne pas être trop similaires à ceux utilisés pour trouver l'hyperplan mais être tout de même situés franchement d'un côté ou l'autre de la frontière. L'efficacité des SVM est supérieure à celle de toutes les autres méthodes sur la classification de textes. Son efficacité est aussi très bonne pour la reconnaissance de formes. Un autre intérêt est la sélection de Vecteurs Supports qui représentent les vecteurs discriminant grâce auxquels est déterminé l'hyperplan. Les exemples utilisés lors de la recherche de l'hyperplan ne sont alors plus utiles et seuls ces vecteurs supports sont utilisés pour classer un nouveau cas. Cela en fait une méthode très rapide.

1.1.6. Programmation génétique :

C'est une méthode générale qui peut être utilisée après n'importe qu'elle méthode précédente, par exemple avec les arbres de décisions. En entrée, un algorithme génétique reçoit une population de classes non optimales. Le but du programme génétique est de produire une classe plus optimale que chacun de ceux de la population d'origine. D'une façon simple, cela consiste à extraire les meilleures parties de chaque classe d'origine et de les mettre ensemble pour produire une nouvelle classe. Cela suppose de pouvoir comparer l'efficacité d'une classe. Un résultat important de la méthode est qu'après chaque itération on obtient une classe meilleure qu'avant. On peut donc arrêter les itérations à tout moment, même si le résultat n'est pas l'optimum.

2. La classification non- supervisée : La classification non-supervisée est utilisée lorsque l'on possède des documents qui ne sont pas classés et dont on ne connaît pas de classification. A la fin du processus de classification non-supervisée, les documents doivent appartenir à l'une des classes générées par la classification.

2.1. Méthodes d'apprentissage non-supervisé :

L'objectif d'une méthode de classification c'est la recherche d'une partition, ou répartition des individus en *classes*, ou catégories. Ceci est fait en optimisant un *critère* visant à regrouper les individus dans des classes, chacune le plus homogène possible et, entre elles, les plus distinctes possible. Cet objectif est à distinguer des procédures de discrimination, ou encore de classement pour lesquelles une typologie est *a priori* connue, au moins pour un échantillon d'apprentissage.

On distingue trois catégories de classifications non-supervisées :

2.1.1. Classification ascendante hiérarchique, ou CAH :

Il s'agit de regrouper itérativement les individus, en commençant par le bas (les deux plus proches) et en construisant progressivement un arbre, ou *dendrogramme*, regroupant finalement tous les individus en une seule classe. Ceci suppose de savoir calculer, à chaque étape ou regroupement, la distance entre un individu et un groupe ainsi que celle entre deux groupes.

Au départ, chaque individu est dans un groupe distinct. A chaque étape, deux groupes sont rassemblés en un seul.il faut un critère d'agrégation.

Exemple :

A-B-C-D-E-F

A-BC-D-E-F

A-BC-DE-F

ABC-DE-F

ABC-DEF

ABCD

On voit bien que à chaque itération, la paire d'objets de classes différentes les plus proches est choisie, et leurs classes sont fusionnées.

2.1.2. Classification descendante hiérarchique :

Au départ tous les individus sont de même groupe. A chaque étape, un groupe est séparé en deux .il faut un critère de séparation.

2.1.3. Classification non hiérarchiques (centres mobiles) :

Le plus connu de ces algorithmes est nommé k-means. Cet algorithme suppose que nous connaissions le nombre de classe voulu k.

Ces algorithmes requièrent une représentation vectorielle des individus munis d'une métrique, généralement euclidienne. Il est important de noter que, contrairement à la méthode hiérarchique précédente, le nombre de classes *k* doit être déterminé *a priori*.

C'est une méthode itérative : après une initialisation des centres consistant, par exemple, à tirer aléatoirement *k* individus, l'algorithme répète deux opérations jusqu'à la convergence d'un critère :
1. Chaque individu est affecté à la *classe* dont le centre est le plus proche au sens d'une métrique.
2. Calcul des *k centres* des classes ainsi constituées.

Pour mettre en œuvre des méthodes de classification il faut faire un choix d'un mode de représentation des documents [Seb., 2002], car il n'existe actuellement aucune méthode d'apprentissage capable de représenter directement des données non structurées (textes). Ensuite, il est nécessaire de choisir une mesure de similarité et enfin, de choisir un algorithme de classification non supervisée.

Les techniques mises en œuvre pour calculer les similarités varient bien évidemment selon les disciplines,

Mais elles s'intègrent cependant le plus souvent dans une même approche générale en deux temps :

1. Les documents textuels sont d'abord associés à des représentations spécifiques qui vont servir de base au calcul des similarités. Bien que la nature précise des représentations utilisées dépende fortement du domaine d'application, il faut noter que, presque dans tous les cas, les documents sont représentés sous la forme d'éléments d'un espace vectoriel de grande dimension.

2. Un modèle mathématique est choisi pour mesurer les similarités

V. la similarité entre document ou entre document/requête :

La similarité sémantique est un concept selon lequel un ensemble de documents ou de termes se voient attribuer une métrique basée sur la ressemblance de leur signification / contenu sémantique. Concrètement, cela peut être réalisé en définissant une similitude topologique, par exemple, en utilisant des ontologies pour définir une distance entre les mots, ou en définissant une similitude statistique, par exemple en utilisant un modèle d'espace vectoriel pour corréler les termes et les contextes à partir d'un corpus de texte approprié (co-occurrence).

Deux concepts sont considérés comme sémantiquement similaires s'il y a une synonymie, hyponymie, antonymie, ou troponymie entre eux (Exemples : MEDECIN-CHIRURGIEN, SOMBRE-CLAIR). Deux sens de mots sont considérés comme sémantiquement liés s'il

existe au moins une relation lexico-sémantique entre eux - classique ou non classique (Exemples : CHIRURGIEN-SCALPEL, ARBRE-OMBRE) [Mohammad and Hirst, 2012].

V.1. Mesures de similarité entre document ou entre document/requête existantes :

Une mesure de similarité est, en général, une fonction qui quantie le rapport entre deux objets, comparés en fonction de leurs points de ressemblance et de dissemblance. Les deux objets comparés sont, bien entendu, de même type.

V.1.1. Métriques :

Toutes les mesures de similarité ne sont pas des métriques. Pour être une métrique, une mesure d doit satisfaire les 4 conditions suivantes :

Soit x, y et z, trois éléments d'un ensemble, et soit d(x, y) la distance entre x et y.

- Positivité : $d(x, y) \geq 0$.
- Principe d'identité des indiscernables : $d(x, y) = 0 \equiv x = y$.
- Symétrie : $d(x, y) = d(y, x)$.
- Inégalité triangulaire : $d(x, z) \leq d(x, y) + d(y, z)$

V.1.2. Similarité Cosinus :

Distances entre vecteurs Ti et Tj dans espace multidimensionnel est :

$$Cos(T_i, T_j) = \frac{T_i . T_j}{\| T_i \| . \| T_j \|}$$

Où Ti.Tj représente le produit scalaire des vecteurs Ti et Tj
|| Ti || et || Tj || représentent respectivement les normes de Ti et Tj
La matrice de similarité est une matrice symétrique de dimension N x N, où N est le nombre de documents à classifier, de diagonale nulle (pour les distances euclidiennes et Manhattan) et de diagonale égale à 1 (pour la distance cosinus), et dont les indices représentent les numéro (index) des documents du corpus à classifier.

V.1.3. Coefficient de corrélation de Pearson :

Le coefficient de corrélation de Pearson calcule la similarité entre deux documents d_1 et d_2 comme le cosinus de l'angle entre leurs représentations vectorielles centrées-réduites. La similarité obtenue $Sim_{Pearson}(d_1, d_2) \in [-1, 1]$.

$$sim_{pearson}(d_1, d_2) = sim_{cosinus}(d_1 - \overline{d_1}, d_2 - \overline{d_2})$$

où $\overline{d_1}$ (resp. $\overline{d_2}$) représente la moyenne de d_1 (resp. d_2).

V.1.4. Distance euclidienne :

La distance euclidienne calcule la similarité entre deux documents d_1 et d_2 comme la distance entre leurs représentations vectorielles ramenées à un seul point.

$$sim_{euclidienne}(d_1, d_2) = \| \vec{d_1} - \vec{d_2} \| = \sqrt{\sum_{i=1}^{n} (d_{1_i} - d_{2_i})^2}$$

V.1.5. Distance (d'édition) de Levenshtein :

La distance de Levenshtein [Levenshtein., 1966] calcule la similarité entre les représentations sous forme de chaines de caractères des documents d_1 et d_2. Il s'agit du cout minimal, i.e. du nombre minimal d'opérations d'édition, pour transformer d_1 en d_2. Les opérations sont les suivantes :

- substitution d'un caractère de d_1 en un caractère de d_2,
- ajout dans d_1 d'un caractère de d_2,
- suppression d'un caractère de d_1.

Pour obtenir la distance de Levenshtein $Sim_{levenshtein}$ (d_1, d_2) entre les documents d_1 et d_2, il s'agit d'associer à chacune de ces opérations un cout. Le cout des opérations est toujours égal à 1, sauf dans le cas d'une substitution de caractères identiques. Notons que cette distance a été étendue pour prendre en compte la grammaire, la phonétique, ...

Les techniques basées sur le modèle vectoriel sont faciles à développer, il s'agit uniquement de calcul vectoriel. Quelque soit la technique utilisée, basée sur le modèle vectoriel, a de fait, le même format initial, à savoir, la représentation vectorielle .mais, Des mots identiques considèrent comme peu pertinents peuvent parfois trop influer sur la valeur de la similarité. Par exemple, pour les phrases "Tout est bien qui finit bien" et "C'est notre seul bien", le terme "est" n'est pas vraiment pertinent et pourtant, il va avoir un poids certain.

Notons cependant que la lemmatisation, l'élimination des mots-vides et le tf-idf permettent de pallier cet inconvénient.

Les techniques basées sur l'approche syntaxique ne laissent pas de place aux exceptions ;Elles sont donc facilement automatisables. Mais, Par définition, les techniques basées sur cette approche ne prennent pas en compte la sémantique. Par exemple, il est difficile de trouver une forte similarité entre "Je possède un chien" et "J'ai un animal".

Dans le contexte de notre étude, plusieurs mots peuvent Être utilisés pour parler du même objet. Par conséquent, la prise en compte de la sémantique semble importante. Les relations syntaxiques sont ignorées. Par exemple, aucune différence n'est faite entre "Pierre aime Marie" et "Marie aime Pierre". Dans le contexte de notre étude, les relations syntaxiques peuvent influer sur la pertinence. Par conséquent, il faudrait trouver un moyen d'incorporer des techniques d'analyse de variation du texte. De même, les rôles sémantiques sont ignorés.

Par exemple, dans "La société A achète la société B" et "La société B a été achetée par la société A", seule la forme verbale change. Cela peut engendrer des problèmes de pertinence. Une proposition serait peut-être d'analyser les classes verbales, sans oublier les problèmes liés aux négations (par exemple, "Je suis malade" et "Je ne suis pas malade") et aux antinomies semblent encore difficiles à pallier.

V.2. Mesures de similarité entre concepts :

V.2.1. Les approches basées sur les arcs (distances) :

1. Les Mesures de Rada & al :

[Rada et al., 1989] suggèrent que, pour mesurer la distance entre deux concepts ontologiques, notée *dist (c1, c2)*, on se base sur le nombre d'arcs minimum à parcourir pour aller du concept *c1* au concept *c2*. La mesure de similarité est ainsi de la forme :

Sim $_{Rada}$ (c1, c2) = 1 / 1+dist (c1, c2) avec : $dist\ (c1, c2) = Min_{chemin}\ (c1, c2)$.

2. La Mesure de Resnik [Resnik, 1995] :

Cette mesure utilise la notion de distance sémantique de la manière suivante : deux concepts sont d'autant plus similaires que la valeur de la distance sémantique entre eux est faible. La similarité est définie par rapport à la longueur des chemins qui relient deux concepts dans la hiérarchie. La similarité entre c1 et c2 est: **(C_1,C_2)=2D-(C_1,C_2)**

Où D est le maximum des longueurs des chemins possibles qui relient c1 et c2 et len(c1 , c2) le plus petit chemin entre c1 et c2 .

3. La Mesure de Hirst-St.Onge. [Hirst et al., 1998] :

L'idée de base de cette mesure est que si deux concepts sont reliés entre eux par un chemin très court et qui « ne change pas la direction » alors les deux concepts sont similaires. Le calcul de la similarité est basé sur le poids du chemin le plus court qui mène d'un concept à un autre et le nombre de changements de directions.

$$Sim_{hirst\text{-}St.}(C1,C2) = T\text{- chemin} - K * D$$

4. La Mesure de Wu-Palmer [Wup., 1994] :

La mesure de similarité de [Wup., 94] est définie comme étant la distance qui sépare deux concepts par rapport à leur concept le plus spécifique (PPG) qui subsume les deux concepts dans l'ontologie ainsi que la racine de la hiérarchie.

Etant donnée une ontologie W formée par un ensemble de nœuds et un nœud racine R .La similarité entre c1 et c2 (voir figure) est :

$$\mathbf{sim_{Wu\_Palmer}}(c_1 , c_2) = \frac{2 * N_3}{N_1 + N_2 + 2 * N_3}$$

Plus formellement cette mesure devient :

$$\mathbf{sim_{Wu\_Palmer}}(c_1, c_2) = \frac{2 * prof(c)}{\mathbf{dist}(c_1, c) + \mathbf{dist}(c_2, c) + 2 * \mathbf{prof}(c)}$$

Où « c » est le concept le plus spécifique qui subsume les deux concepts c1 et c2, prof (c) est le nombre d'arcs qui sépare c de la racine et dist(ci, c) est le nombre d'arcs qui séparent ci de c.

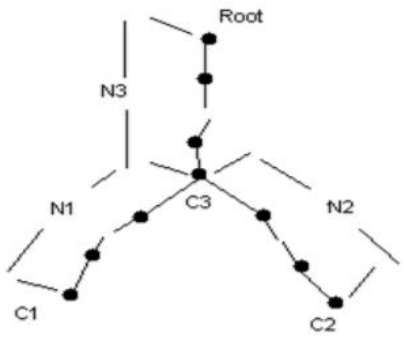


Figure 7 : Les Relations Conceptuelles (Wu & Palmer, 1994)

La mesure de [Wup., 1994] est intéressante mais présente une limite car elle vise essentiellement à détecter la similarité entre deux concepts par rapport à leur distance de leur plus petit généralisant, ce qui ne permet pas de capter les mêmes similarités que la similarité conceptuelle symbolique. Cependant avec cette mesure on peut obtenir une similarité plus élevée entre un concept et son voisinage par rapport à ce même concept et un concept fils, ce qui est inadéquat dans le cadre de la recherche de l'information où il faut ramener tous les fils d'un concept (i.e requête) avant son voisinage.

5. La mesure de Leacock et Chorodow [Leacock et al., 1998] :

Cette mesure est basée sur le plus court chemin entre deux concepts.

$$sim_{\ Leacok_{chorodow}}(c_1, c_2) = -Log \frac{\min Len\ (c_1, c_2)}{2 * D}$$

Où min Len (C_1,C_2) est la longueur du plus court chemin entre c1 et c2 et D est la profondeur maximale de l'ontologie.

6. La Mesure de Zargayouna :

Dans cette mesure [Zargayouna et al., 2004] ont définies une fonction *spec(c1, c2)* qui calcule la spécificité de deux concepts par rapport au concept le plus bas de l'ontologie (Bottom, concept virtuel qui symbolise la fin de l'ontologie) comme le montre la figure .

Cette fonction servira à pénaliser les concepts qui ne sont pas dans la même lignée. Ainsi elle assure que les fils sont pris en compte en priorité et qu'aucun concept du voisinage ne sera plus similaire que les fils.

spec(c1, c2) = N4*N1*N2. (Voir figure) Plus formellement :

$$\mathbf{spec(C1,C2)=prof_b(C) * dist(C2\ ,C).}$$

Avec $prof_b(c)$ correspond au nombre maximum d'arcs qui séparent le plus petit ancêtre commun du concept virtuel représentant l'anti-racine de l'ontologie.

Ainsi la mesure de similarité de [Wup., 19 94] devient :

$$sim_{Zargayouna}(c_1, c_2) = \frac{2 * prof(c)}{dist(c_1, c) + dist(c_2, c) + 2 * prof(c) + spec(c_1, c_2)}$$

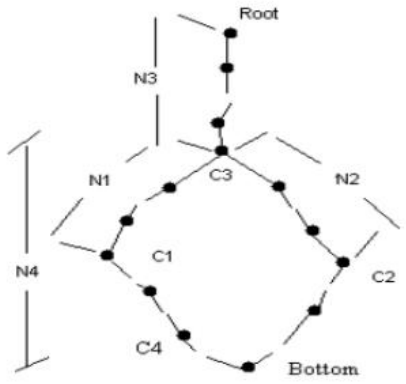


Figure 8 : Les Relations Conceptuelles (Zargayouna et al., 2004)

V.2.2. METHODE BASEE SUR LE CONTENU INFORMATIF (NŒUDS) :

Les mesures utilisées dans ces approches sont fondées sur la notion de contenu informationnel (informatif) qui utilise conjointement l'ontologie et le corpus. Le contenu informatif d'un concept traduit la pertinence d'un concept dans le corpus en tenant compte de sa spécificité ou de sa généralité. Pour se faire la fréquence de concepts dans le corpus est calculée pour retrouver le contenu informatif. Cette fréquence regroupe la fréquence d'apparition du concept lui-même ainsi que les concepts qu'il subsume (concepts fils). Les mesures les plus connus dans cette catégorie sont celles de [Resnik., 1999], [Lin., 1998], [Jiac., 1997].

1. La Mesure de Resnik [Resnik.,1999] :

[Resnik., 19 99] a défini la similarité sémantique entre 02 concepts par la quantité d'informations qu'ils partagent en commun et elle est égale au contenu informationnel du concept le plus spécifique (ppg : plus petit généralisant) qui subsume les deux concepts dans l'ontologie et elle est définie comme suit :

$$(C_1 , C_2) = CI\ (ppg(C_1 , C_2)).$$

Avec CI=-log (P(c)), où P(c) est définie comme la probabilité de retrouver un mot du corpus qui soit une instance du concept c:

$$P(c) = \frac{Freq(c)}{N}$$

Tel que N est la taille totale d'échantillon de texte et Freq(c) est la fréquence d'occurrence des mots dénotant le concept c dans la collection.

2. La Mesure de Lin [Lin ., 1998] :

La similarité entre deux concepts est mesurée par le ratio du contenu d'information nécessaire pour mesurer la « communalité » des deux concepts, sur le montant du contenu d'information nécessaire pour décrire chacun des deux concepts. La similarité entre deux concepts c1 et c2 est :

$$sim_{Lin}(c_1, c_2) = \frac{2 * CI(psc(c_1 , c_2))}{CI(c_1) + CI(c_2)}$$

V.2.3. METHODES HYBRIDES :

Ces méthodes sont fondées sur un modèle mixte qui combine entre des approches basées sur le comptage des liens en plus du contenu informatif qui est considéré comme facteur de décision.

1. La Mesure de Jiang et Corath [Jiang et al., 1997] :

La similarité est définie comme une distance sémantique qui tient compte aussi des contenus d'informatifs dans la fonction de la similarité. La distance sémantique est calculée comme suit : **distance $(C_1 , C_2) = 2 * CI(psc(C_1 , C_2)) - (CI(C_1) + CI(C_2))$**

$$sim_{Jiang\_Corath}\ (c_1 , c_2) = \frac{1}{distance(c_1, c_2)}$$

2. LA Mesure de Leacock et Chodorow :

Une autre méthode présentée par [Leacock et al., 1998] qui combine entre la méthode de comptage des arcs et la méthode du contenu informationnel. La mesure proposée par Leacock et Chodorow [Leacock et al., 1998] est basée sur la longueur du plus court chemin entre deux synsets de Wordnet. Les auteurs ont limité leur attention à des liens hiérarchiques «is-a » ainsi que la longueur de chemin par la profondeur globale de la taxonomie. La formule est définie par :

$$sim_{Lec}(c_1, c_2) = -log \frac{cd(c_1, c_2)}{2 * M}$$

Où M est la longueur du chemin le plus long qui sépare le concept racine de l'ontologie, du concept le plus en bas. On dénote par cd (X, Y) la longueur du chemin le plus court qui sépare X de Y.

Conclusion :

Dans ce chapitre nous avons parlé de similarité et de son utilisation dans des différents domaines, on a spécifié la recherche d'information, la détection plagiat et la classification des documents. On s'est basé sur les différent approches de classification des documents, où on a montré qu'il Il existe une quantité importante de méthodes qui sont toutes issues des recherches basé sur l'apprentissage. On a aussi vu que l'étape de représentation des documents est essentielle, Et que la plupart des méthodes nécessitent de représenter chaque document sous forme d'un vecteur.

En fin, nous avons classifié les approches pour l'identification de la similarité syntaxique et sémantique. Nous avons cité Les approches basées sur les nœuds utilisant des mesures du contenu informationnel pour déterminer la similarité conceptuelle. L'autre famille d'approche basée sur les distances des arcs qui est basé sur le plus court chemin entre les nœuds. Et l'approche hybride qui combine entre les deux premières approches.

Chapitres IV : Réalisation

Introduction :

Au cours du chapitre précédent nous avons donné un aperçu des méthodes les plus utilisées dans le calcul de la similarité sémantique des documents textuels. Dans ce chapitre nous allons parler de l'implémentation de l'approche de similarité sémantique qu'on a choisie.

I. la méthode implémentée :

Nous avons retenu le calcul de similarité défini par la méthodeProxigénéa .Nous avons simplifié la formule de calcul comme le montre l'équation (1).

Après projection du contenu d'un document sur WordNet, nous obtenons pour chaque document un ensemble de concepts: (annotations = concepts et relations entre concepts, extraits de WordNet, et qui correspondent au contenu d'un document).

La similarité entre deux annotations correspondant à deux documents est définie comme la moyenne pondérée des similarités entre les concepts qui les composent.

Soit : - A1,A2 deux annotation ,qui sont composés d'un ensemble de concept.

- Concepts (A) :l'ensemble des concepts de A.

- C1i et C2j des concepts appartenant respectivement aux annotations A1 et A2.

- $coef\,(Ci)$ la fonction déterminant le degré de d'importance de concept d'une annotation A, dans notre travail on a défini *coef(Ci)=1*.

- $simconcept\ (C1i, C2j)$ est la similarité entre les concepts C1i et C2j.

$$Sim(A1, A2) = \frac{\sum_{i=1}^{/Concepts(A1)/} coef(C1i).\ \underset{j=1}{\overset{/Concepts(A2)/}{Max}}\ (simconcept(C1i, C2j))}{\sum_{i=1}^{/Concepts(A1)/} coef(C1i)} \qquad (1)$$

Dans notre travail la *simconcept(C1i, C2j)*est déterminé par la mesure de similarité proposé par Wu Palmer.

Exemple :

Soient deux documents d1 et d2.

d1-------A1 ={C1,C2,C3}

d2 ------A2={C4,C5}

On calcule

1/ Max1 = Max (sim(C1,C4), sim(C1,C5))

Max2 = Max (sim(C2,C4), sim(C2,C5))

Max3 = Max (sim(C3,C4), sim(C3,C5))

2/ Sim (A1,A2) = (Max1+Max2+Max3)/3.

II. la collection des documents utilisés : Nous avons utilisé une collection de 10 documents textuels extraits depuis wikipedia, pour calculer la similarité entre ces documents.

Numéro de document	**Nom de document**	**Domaine**	**Résumé**
01	Chirurgie	Médecine	Le document présente la définition de la chirurgie et ces différentes spécialités existantes, et la durée de formation en chirurgie.
02	Ebola	Médecine	Le document expose la maladie d'Ebola et ces symptômes, et la façon pour le reconnaitre, et le traitement utilisé.
03	Infection	Médecine	Le document présente la maladie d'infection, ces causes, les méthodes de lutter contre cette maladie et le traitement utilisé.
04	Tumeur	Médecine	Le document présente une petite introduction sur Néoplasique et ces quatre types.
05	Zika	Médecine	Le document consiste la virus Zika, ces symptômes, sa propagation.
06	Médecine	Médecine	On trouve dans le document l'historique de terme médecine.
07	Programming-language	Informatique	Le document parle sur le langage de programmation.
08	Laptop	Informatique	Le document présente le laptop et ces composantes.
09	Desktop-computer	Informatique	Le document présente l'ordinateur de bureau et ces composantes.
10	Wireless_network	Informatique	Le document présente une définition de réseau sans fil.

III. Description du l'environnement technologique :

Pour le développement de cette application nous avons eu recours à plusieurs composantes technologiques et API, on citera :

- **NetBeans [1]:** nous avons utilisé l'IDE NetBeans comme Environnement de développement.
- **WordNet 2.1[2] :** qui est une base de données développé par l'université de Princeton a la quel on peut accéder à partir des interfaces pour de nombreux langages de programmationet nous l'avons utilisé comme ressource pour extraireles sens des mots.
- **JWNL API [3]:** JWNL est une API pour accéder à des dictionnaires relationnels WordNet style. Il fournit également des fonctionnalités au-delà de l'accès aux données, telles que la relation découverte et traitement morphologique. JWNL est une implémentation de Java pur de l'API WordNet, ce qui signifie tout ce qui est nécessaire est les bibliothèques Java et les fichiers de dictionnaire, nous avons utilisé la version 1.4 qui est compatible avec WordNet 2.1.
- **RiTa et RiWordnet API[4] :** Fournit un appui pour l'accès à la base de données WordNet 2.1.

[1] :https://www.oracle.com/fr/index.html
[2] :https://wordnet.princeton.edu/
[3] :http://jwordnet.sourceforge.net/handbook.html
[4] :https://rednoise.org/rita/reference/RiWordNet.html

IV. Description de processus de notre application :

Notre application permet à travers plusieurs étape de calculer la similarité entre deux annotation, le figure 9 montre une vue d'ensemble de notre processus.

Etape 1 : Extraction des Noms :

Cette étape consiste à parcourir les deux textes et extraire tous les noms.

Etape 2 : Récupération des bases des mots :

Après avoir récupéré tous les noms pour chaque texte on récupèrera ensuite pour chaque nom sa base en utilisant quoi ?, et on récupérera ensuite le sens de ces noms en exploitant Wordnet.

Etape 3 : Projection et récupération des sens à partir de WordNet :

Cette étape consiste à projeter les différents noms extraits d'un document sur Wordnet pour récupérer le sens adéquat de chaque terme par un processus de désambigüisation.

Etape 4 : désambiguïsation des termes :

Dans cette étape le but est de désambiguïser les termes pour choisir le sens le plus adéquat et le plus précis pour chaque nom. On a choisit la désambiguïsation **naïve** qui consiste à retourner le premier synset de Wordnet correspondant à un nom donné.

Etape 5 : Calcule de similarité :

Apres avoir désambiguïsé les noms, nous calculons la similarité entre ces deux texte.

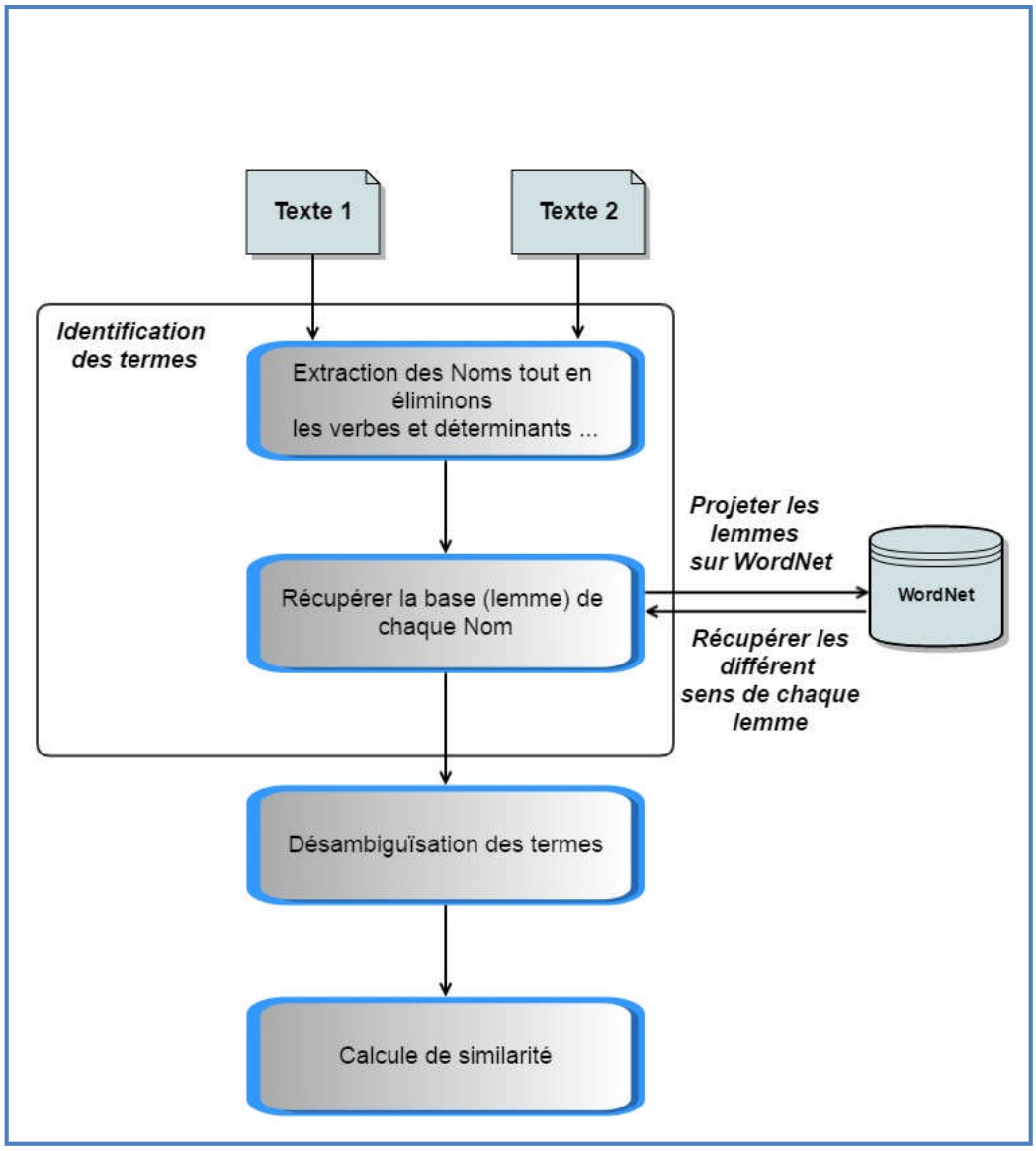


Figure 9 : Etapes de calcul de la similarité sémantique de deux documents

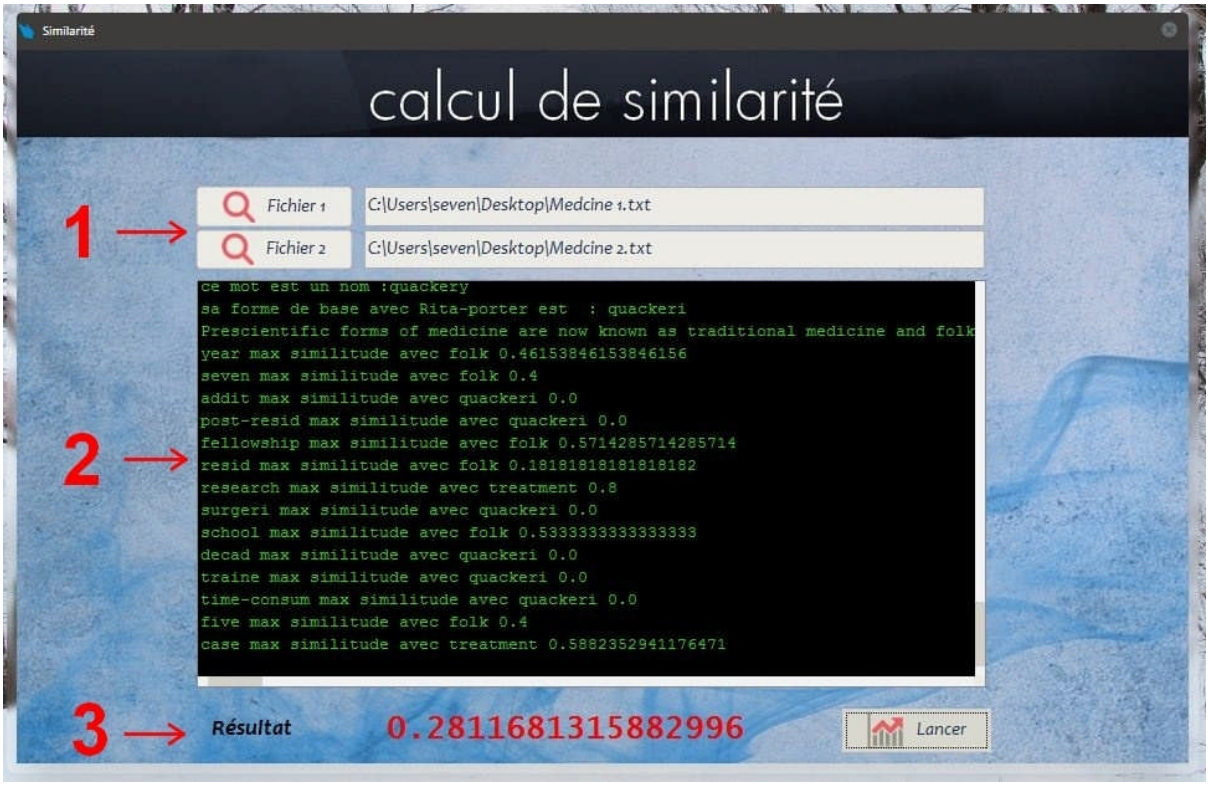
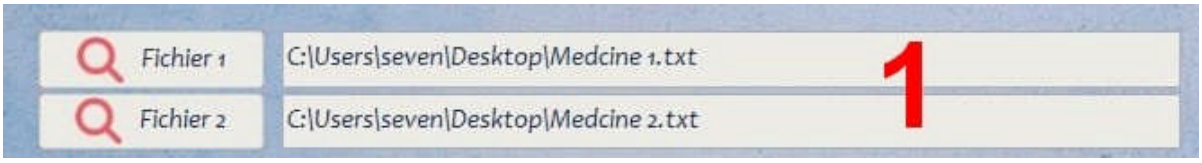
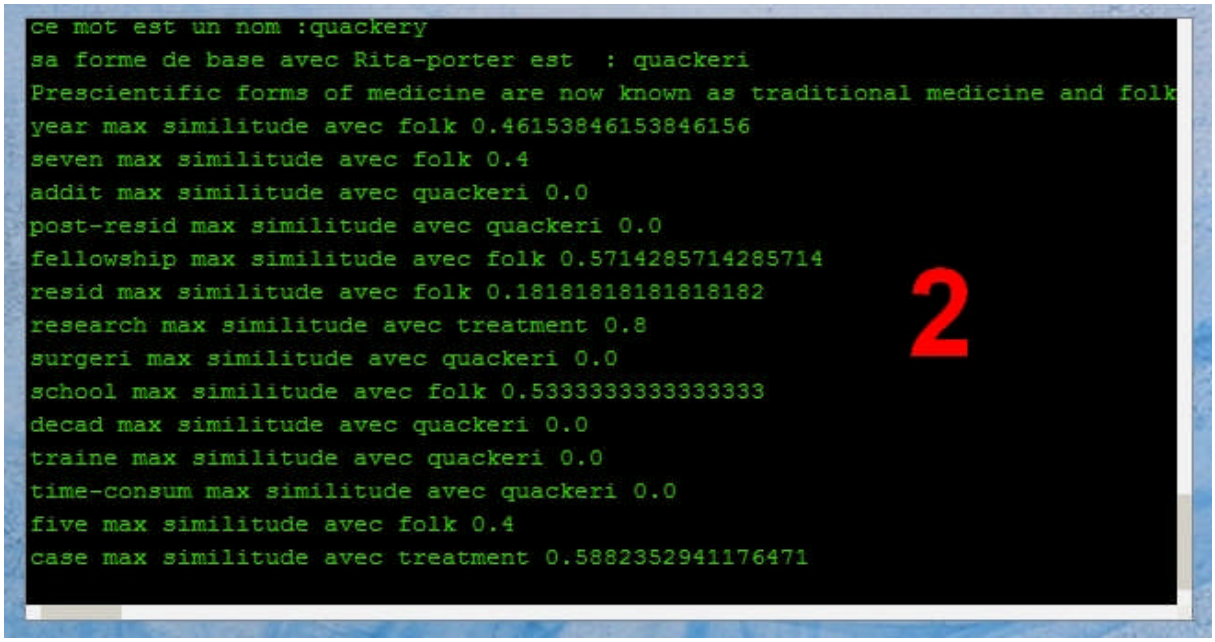


Figure 10 : Fenêtre principale de l'application

- Comme on peut le remarquer l'application a une interface simple et facile à utiliser.



1- Au lancement, l'application prend en entrée deux fichiers Textes (.txt).



2- Dans cette partie de l'interface on peut constater et suivre en temps réel les traitements.

V. Les résultats obtenus sur la collection : le tableau montre les résultat de similarité entre les documents de notre collection comparés deux à deux.

Document	**01**	**02**	**03**	**04**	**05**	**06**	**07**	**08**	**09**	**10**
01	/	0.857	0.631	0.615	0.666	0.571	0.028	0.039	0.012	0.018
02	0.857	/	0.888	0.769	0.875	0.625	0.014	0.025	0.090	0.004
03	0.631	0.888	/	0.588	0.875	0.588	0.027	0.0138	0.007	0.009
04	0.615	0.769	0.588	/	0.625	0.636	0.033	0.024	0.0116	0.0246
05	0.666	0.875	0.875	0.625	/	0.555	0.0051	0.005	0.007	0.003
06	0.571	0.625	0.588	0.636	0.555	/	0.015	0.022	0.0051	0.008
07	0.028	0.014	0.027	0.033	0.0051	0.015	/	0.875	0.666	0.355
08	0.039	0.025	0.0138	0.024	0.005	0.022	0.875	/	0.705	0.875
09	0.012	0.090	0.007	0.0116	0.007	0.0051	0.666	0.705	/	0.281
10	0.018	0.004	0.009	0.0246	0.003	0.008	0.355	0.875	0.281	/

Tableau 02 : les résultats de similarité entre les documents de notre collection.

Analyse des résultats :

Afin d'évaluer le calcul de la similarité sémantique entre documents, nous avons testé notre application sur une collection de 10 documents. Le tableau ci-dessus, Montre les résultats obtenus. La valeur de la similarité entre documents varie entre 0 et 1.

- Les documents qui appartiennent au même domaine ont une similarité proche de 1, car ils partagent des concepts en commun. Se sont tous des documents qui appartiennent au domaine de médecine. Nous donnons ci dessous

Document	Exemples de synsets communs	ExempesSynsets différents	Sim (D1,D2)
D1 (02) Ebola	Tissu, health, word, cancer,laboratori, symptom, traitement, death, cholera, control,	Date, individu, diagnosi, mening, number, but, contact, west, simple, fluid, epidem.	0.875
D2 (04) Tumeur	Tissu, health, word, cancer, laboratori, symptom, traitement, Death, cholera, control,	Mass, creation, pattern, organiz, cell, plasma, form, growth, dysplasia.	

Document	ExemplesSynsets communs	ExempesSynsets différents	Sim (D1,D2)
D1 (01) chirurgie	Year, research, fellowship, spread, virus, seven, resid,	Addit, surgeri ,decad, traine,	0.857
D2 (05) zika	Year, research, fellowship, spread, virus, ten, dna,	Dominican, vaccine,dose,fda, mice.	

Document	ExemplesSynsets communs	ExempesSynsets différents	Sim (D1,D2)
D1 (08) laptop	Use, specif, field, comput, form, construct, model, element, market.	Medel, featur , screen, processor, unit,speaker, organ.	0.875
D2 (07) programming-language	Use,specif, field, comput, form, program, behavior.	Age, piano, musa, year, playear, result.	

Document	ExempesSynsets communs	ExempesSynsets différents	Sim (D1,D2)
D1 (09) Desktop-computer	Comput, unit, desktop, case, input, screen, display, screen, capabl, touchpad, processor, memori, component.	All-in-on, AC, adapt, power, organ, varieti , Purpose.	0.705.
D2 (10) laptop	Comput, unit, desktop, factor, output, keyboard, display, screen, capabl, touchpad, processor, memori, component.	Account, webcam, batteri, notebook, macbook.	

- Les documents qui appartiennent aux différents domaines ont une similarité proche de 0. Car la majorité de leurs concepts sont différents Comme :

Document	ExempesSynsets communs	ExempesSynsets différents	Sim (D1,D2)
D1 (03) Infection	System, infect, bacteria, branch,	Microparasit, death, antifung, virus, medicin,	0.009.
D2 (10) Wireless	Implement, connect, layer, network,	Satellite, process, sattelit,place.	

Document	ExempesSynsets communs	ExempesSynsets différents	Sim (D1,D2)
D1 (05) Zika	Approach,ten, market. Adult,virus.	Prioriti,dangu, defect, Symptom, dose, vaccin, world, infect, center,spread.	0.005.
D2 (08) laptop	Design, unit, work, child,virus.	Account, power, construct, AC ,apple, processor, use,field, factor.	

Document	ExempesSynsets communs	ExempesSynsets différents	Sim (D1,D2)
D1 (09) Desktop_computer	Unit, connect, cabl, equipement, instal.	Comput, all_in_on, phone, cell, osi, implement,	0.281.
D2 (10) Wireless_network	Level, connect, cabl, equipement, instal.	Sattelit, processor,network, structur,	

L'analyse des résultats résumés dans le tableau 02 permet d'affirmer que le calcul de similarité utilisé donne des résultats probants.

Conclusion :

Ce chapitre a été dédié à la présentation de l'application, l'environnement technique de développement de notre application et aussi les différentes ressources utilisées pour la réalisation de cette application, ainsi que les résultats de l'expérimentation de l'application.

Conclusion générale

Dans ce mémoire, Nous avons exposé le problème d'augmentation de volumes des documents textuels et nous avons évoqué la nécessité de développer des techniques permettant une utilisation ciblée et efficaces de ces données en utilisant des mesures de similarité sémantique.

Nous avons orienté notre travail sur le calcul de la similarité des documents textuels.

La notion de similarités entre documents textuels constitue un domaine actif de recherche. La similarité sémantique est utilisée dans le domaine de recherche d'information, le domaine de détection de plagiat, et le domaine classification des documents. Par exemple, en Recherche d'information, les documents pertinents retournés par le moteur de recherche sont les plus proches de la requête selon une certaine mesure de similarité, de même, dans le cas de classification des documents, les documents sont également regroupés en classes en fonction d'une mesure de similarité spécifique.

Nous avons implémenté une formule permettant de calculer la similarité entre documents textuels .Ce calcul est basé sur l'utilisation d'une ressource externe représenté par Wordnet. Nous avons effectué des tests sur un petit corpus, les résultats obtenus sont probants. Néanmoins, l'utilisation d'un corpus plus large et un nombre de domaine plus grand permettra de conclure sur l'efficacité de l'approche retenue.

Une comparaison avec des calculs de similarité classique (cosinus…) permettra de juger sur l'impact de la prise en compte de la sémantique dans le calcul de la similarité entre documents textuels.

Bibliographie

[ABBAS., 2014] ABBAS Nacira 2014. Vers une Extension Sémantique de l'Analyse Formelle de Concepts : Application à la Recherche d'Informations. Mémoire de magister en Informatique, UNIVERSITE MOULOUD MAMMERI, TIZI-OUZOU. (2014), p32-33

[Abi Chahine., 2011] : Indexation et recherche conceptuelles de documents pédagogiques guidées par la structure de Wikipédia. Thèse de Doctorat en Informatique. L'Institut National des Sciences Appliquées de Rouen, Octobre 2011.

[Aas&Eik. , 1999] K. Aas and L. Eikvil. Text Categorization : A Survey. Technical report. Norwegian Computing Center. 1999.

[Apt&al., 1994] C. Apté, F.J. Damerau, and S.M. Weiss. Automated Learning of Decision Rules for Text Categorization. ACM Transactions on Information Systems. 1994. 12(3) : pp 233-251.

[Baeza., 2011]Baeza-Yates, R., Ribeiro-Neto, B. A. Modern Information Retrieval.Pearson Education Ltd., Harlow, UK, 2nd edn, 2011.

[Baeza-Yates & al., 1999] : R. Baeza-Yates and R. A. Ribeiro-Neto. Modern information Retrieval. New York : ACM Press ; Harlow England : Addison-Wesley, cop., 1999.

[Baziz., 2005] M. Baziz. "Indexation conceptuelle guidée par ontologie pour la recherche d'information, Thèse de doctorat en informatique", Université Paul Sabatier de Toulouse, 2005.

[Bes&al., 2001] R. Besançon, A. Rozenknop, J.C. Chappelier and M. Rajman. Intégration probabiliste de sens dans la représentation de textes. Actes de la 8ème conférence annuelle sur le Traitement Automatique des Langues Naturelles (TALN). 2001, Tours, France. pp 83-91.

[Bes&al ., 2001] R. Besançon, A. Rozenknop, J.C. Chappelier and M. Rajman. Intégration probabiliste de sens dans la représentation de textes. Actes de la 8ème conférence annuelle sur le Traitement Automatique des Langues Naturelles (TALN). 2001, Tours, France. pp 83-91.

[Boubekeur., 2008] Boubekeur F. "Contribution à la définition de modèles de recherche d'information flexibles basés sur les CP-Nets", thèse de doctorat en informatique, Université Paul Sabatier.(2008).

[Brachman., 1983] R.J. Brachman, "What IS-A is and isn't : An analysis of taxonomic links in semantic networks", in "IEEE Computer", volume 16 no 10, 1983.

[Boucham., 2009] BOUCHAM Souhila Thème Une approche basée Ontologies pour l'indexation automatique et la Recherche d'Information Multilingue (RIM), 2009.

[Buc&al .,1992] C. Buckley, G. Salton and J. Allan. Automatic Retrieval with Locality Information using Smart. Proceedings of the First Text Retrieval Conference. Gaithersburg, 1992. pp 59-72.

[Caro.,1997] Budi Yuwono, Savio L.Y. Lam, Jerry H. Ying, Dik L. Lee, A World Wide Web Ressource Discovery System, 1997.

[Cha., 1990] J. Chauché. Détermination sémantique en analyse structurelle : Une expérience basée sur une définition de distance. TA Information. 1990. Volume. 31, N°1, pp 17-24.

[Charhad., 2005] M. Charhad. "Modèles de Documents Vidéo basés sur le Formalisme des Graphes Conceptuels pour l'Indexation et la Recherche par le Contenu Sémantique". pages 24-25, Novembre 2005.

[Cleverdom., 1970]Cleverdon, C. Progress indocumentation.Evaluation of informationretrievalsystems.Journal of Documentation, 26, pp. 55–67, 1970.

[Cleveland & al., 2000] Cleveland, D. B. ,& Cleveland, A. D. (2000). Introduction to Indexing and Abstracting : (3rded.) Libraries Unlimited.[Cleveland & al., 00] Cleveland, D. B. ,& Cleveland, A. D. (2000). Introduction to Indexing and Abstracting : (3rded.) LibrariesUnlimited.

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

[Declaris., 1994]: N. DeClaris, J. James, A. Nerode, W. Kohn, Intelligent integration of medical models, Proc. IEEE Conference on Systems, Man, and Cybernetics, San Antonio,1994.

[Dee&al ., 1990] S. Deerwester, S. Dumais, T. Landauer, G. Furnas and R. Harshman. Indexing by Latent Semantic Analysis. Journal of the American Society for Information Science. 1990. 41(6): pp 391-407.

[Denos., 1997] Denos, N. « Modélisation de la pertinence en recherche d'information: modèle conceptuel, formalisation et application ». Thèse de Doctorat de l'Université JosephFourier-Grenoble I, 1997.

[Dum&al., 1998] S. Dumais, J. Platt, D. Heckerman and M. Sahami. Inductive Learning Algorithms and Representations for Text Categorization. Proceedings of the 7th ACM International Conference on Information and Knowledge Management (CIKM 1998). ACM Press. 1998, Nex York, USA. pp 148-155.

[Ehrig., 2004] M.Ehrig, P.Haase, M.Hefke et N.Stojanovic. Similarity for ontology-a comprehensive framework. In Workshop Enterprise Modelling and Ontology: Ingredients for Interoperability, 2004.

[Fir .,1957] J. Firth. A Synopsis of Linguistic Theory 1930-1955. Studies in Linguistic Analysis. Réédité, Selected Paper of J.R. FIRTH, F. PALMER (edts.), Longman. 1957. pp 82-95.

[Gangemi et al., 1999] Gangemi A, Pisanelli DM, Steve G: An Overview of the ONIONS Project: Applying Ontologies to the Integration of Medical Terminologies.Data and Knowledge Engineering, 1999, 31, pp. 183-220 (1999).

[Guarino et al., 1999] Guarino, N., C. Masolo, and G. Vetere, OntoSeek: Using Large Linguistic Ontologies for Accessing On-Line Yellow Pages and Product Catalogs, . 1999, National Research Council, LADSEBCNR: Padavo, Italy.

[Gruber., 1993] T. Gruber. "A translation approach to portable ontology specifications.Knowledge .Acquisition". 5(2):199–220, 1993.

[Guarino et al., 2000] Nicola Guarino, Christopher A. Welty: Ontological Analysis of Taxonomic Relationships. ER 2000: 210-224

[Guarino et al., 2002] Nicola Guarino, Christopher A. Welty: Evaluating ontological decisions with OntoClean. Commun. ACM 45(2): 61-65 (2002).

[Har ., 1988] Z. Harris. Language and Information. Columbia University Press. New York, 1988. 120p.

[Har&al., 1989] Z. Harris, M. Gottfried, T. Ryckman, P. Mattick, A. Daladier, T. N. Harris and S. Harris. The Form of Information in Science: Analysis of an Immunology Sublanguage. Preface by H. PUTNAM. Boston studies in the phlisophiy of science. Kluwer Academic Publishers. Boston, USA, 1989. 590p.

[Harter., 1975] Harter, S.A probabilistic approach to automatic keyword indexing.part ii. an algorithm for probabilistic indexing.Journal of the American Society for Information Science (JASIS) 35, 3 (1975), 280–289.

[Harter., 1992] Harter, S. «Psychological relevance and information science », Journal of the American Society for information Science (JASIS),1992.

[Hernandez., 2006] Hernandez N. "Ontologie de domaine pour la modélisation du contexte en recherche d'information", thèse de doctorat en informatique, Université Paul Sabatier. (2006).

[Hirst et al., 1998] Hirst G., and St-Onge D. Lexical chains as representation of context for the detection and correction malapropisms. In Christiane FELLBAUM, editor, WordNet : An electronic lexical database, chapter 13, pages 305–332. The MIT Press.

[Hof., 1999] T. Hofmann. Probabilistic Latent Semantic Analysis. Proceedings of Uncertainity in Artificial Intelligence Conference (UAI 1999). 1999, Stockholm. pp 289-296.

[Ingwersen.,1994] P. Ingwersen. "Polyrepresentation of information needs and semantic entities: elements of a cognitive theory for information retrieval interaction". In Proceedings of the Seventeenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval., pages 101-110, 1994.

[Jai&al ., 2005] S. Jaillet, M. Teisseire et G. Dray. Adéquation des modèles de représentation aux méthodes de catégorisation. Revue Ingénierie des Systèmes d'Information (ISI), numéro spécial « Fouille de données complexes ». 2005. 19p.

[Jiac., 1997] J. Jiang et D. Conrath. Semantic similarity based on corpus statistics and lexical taxonomy. In Proceedings of International Conference on Research in Computational Linguistics, Taiwan, 1997.

[Jiang et al., 1997] Jiang J.J., and Conrath D.W.,. Semantic similarity based on corpus statistics and lexical taxonomy. In International Conference Research on Computational Linguistics (ROCLING X).

[Jérémy & Alain., 2015] Jérémy Ferrero, Alain Simac-Lejeune. D´etection automatique de reformulations - Correspondance de concepts appliqu´ee `a la d´etection du plagiat. 15e Conf´erence Internationale Francophone sur l'Extraction et la Gestion des Connaissances, Jan 2015, Luxembourg, France.p3-5.

[Koh&al ., 2000] T. Kohonen, S. Kaski, K. Lagus, J. Salojärvi, J. Honkela, V. Paatero and A. Saarela. Organization of a Massive Document Collection. IEEE Transactions on Neural Networks. 2000. Volume 11, N° 3. pp 574-585.

[Kompaoré., 2008] N.D.Y. Kompaoré. "Fusion de systèmes et analyse des caractéristiques linguistiques des requêtes: vers un processus de RI adaptatif». Thèse de doctorat en informatique, Université Paul Sabatier de Toulouse, 2008.

[Krovetz., 1997] R. Krovetz. 1997. 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.

[Lancaster., 1998] Lancaster F.-W., 1998, Indexing and abstracting in theory and practice. Library Association Publishing. London.

[Lan&al., 1998] T. Landauer and P. Foltz and D. Laham. Introduction to Latent Semantic Analysis. Discourse Process. 1998. 25: pp 259-284.

[Lan&Lit., 1991] T.K. Landauer, M. Littman. A Statistical Method for Language Independent Representation of the Topical Content of the Text Segments. Actes du 10éme congrès sur l'Information, la Documentation et le Transfert de connaissances (IDT 1991). Ed. Adbs & anrt (edts.).

[Leacock et al., 1998] Leacock C., and Chodorow M. Combining local context and WordNet similarity for word sense identification. In Fellbaum, C., editor, WordNet : an electronic lexical database, volume 11 of Language, Speech and Communication, pages 265–283. The MIT Pr, Cambridge, Massachusetts.

[Lee et al., 1993] Lee J.H., Kim M.H. et Lee Y.J. . Information Retrieval based on conceptual distance in IS-A hierarchies. Journal of Documentation, vol. (49/2) : 188-207.

[Lee et al., 1993]: Joon Ho Lee, Myong Ho Kim, and Yoon Joon Lee. "Information retrieval based on conceptual distance in IS-A hierarchies". Journal of Documentation, 49(2):188-

207, 1993.

[Lee., 1995] J. H. Lee. Combining Multiple Evidence from Different Properties of Weighting Schemes. Proceedings of the 18th Annual International ACM Conference on Research and Development in Information Retrieval (SIGIR 1995). Washington, USA, 1995. pp 180-188.

[Levenshtein., 1966] Levenshtein, V. (1966). Binary codes capable of correcting deletions, insertions, and reversals. Soviet Physics Doklady, 10(8) :707710.

[Lew.,1992b] D. Lewis. Representation and Learning in Information Retrieval. PhD thesis, Department of Computer Science, University of Massachusetts. 1992, USA.

[Lin ., 1998] Lin D. An information-theoretic definition of similarity. In Proc. 15th International Conf. on Learning, Morgan Kaufmann, San Francisco, CA, pp. 296–304 .

[Luhn., 1958] Luhn, H. « The automatic creation of literature abstracts ». IBM Journal of Research and Development 24, 2 (1958), 159–165.

[Maisonnasse., 2008] L. Maisonnasse. “Les supports de vocabulaires pour les systèmes de recherche d’information orientés précision : application aux graphes pour la recherche d’information médicale”. thèse de doctorat en informatique, Université Joseph Fourier- Grenoble I, France, 2008.

[Mezzaro., 1997] .Mizzaro, S. Relevance, the whole (hi) story. Journal of the American Society for Information Science, 48, pp. 810–832, 1997.

[Michelle Bergadaà. ;2008] Commission Ethique-plagiat Université de Genève. LA RELATION ETHIQUE-PLAGIAT DANS LA REALISATION DES TRAVAUX PERSONNELS PAR LES ETUDIANTS. Université de Genève(2008),p112.

[Mizoguchi., 1997] Towards ontology engineering. In The Joint 1997 Pacific Asian Conference on Expert systems - International Conference on Intelligent Systems, p. 259–266, Singapore.

[N. Guarino., 1998] Formal ontology in information systems. In the 1st International Conference on Formal Ontology in Information Systems (FOIS), p. 3–15, Trento, Italy:IOS Press.

[Niles et al., 2003] Niles, I., and Pease, A., (2003). Linking Lexicons and Ontologies: Mapping WordNet to the Suggested Upper Merged Ontology, Proceedings of the IEEE International Conference on Information and Knowledge Engineering, pp 412-416.

[Quillian., 1968] M. Quillian. Semantic Memory. In M. Minsky (Ed.), Semantic information Processing. The MIT Press, Cambridge, MA, 1968. AlsoPhDThesis, Carnegie Institute of Technology, 1967.

[Rada ., 1989] R. Rada, H. Mili, E. Bicknell, and M. Blettner. Development and application of a metric on semantic nets. pages 17–30, 1989.

[Raj&al ., 2000] M. Rajman, R. Besançon et J-C Chappelier. Le modèle DSIR: une approche à base de sémantique distributionnelle pour la recherche documentaire. Traitement automatique des langues, 41(2). 2000. pp 549-578.

[Resnik., 1995] Resnik P. Using information content to evaluate semantic similarity in a taxonomy. In IJCAI, pages 448–453

[Resnik., 1999] Resnik P. Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language, Journal of Artificial Intelligence Research, pages 95–130.

[Robertson., 1976] S.Robertson & K.Sparck Jones, Relevance Weighting for Search Terms. Journal of The American Society for Information Science, Vol 27, N°3, 1976.

[Roussey& al., 2001] Roussey, C., Calabretto, S. etPinon, J.-M. (2001a). A multilingual information system based on knowledge representation. pages 98-111.

[Sal ., 1971a] G. Salton. The SMART Retrieval System. Experiments in Automatic Document Processing. Prentice Hall, Englewood Cliffs. 1971.
[Sal ., 1971b] G. Salton. The Performance of Interactive Information Retrieval. Information Processing Letters. 1971. (2): pp 35-41.
[Sal., 1973] G. Salton. Recent Studies in Automatic Text Analysis and Document Retrieval. ACM Journal. 1973. 20(2) : pp 258-278.
[Salton et al., 1983] G. Salton, E.A. Fox, H. Wu. *Extended Boolean information retrieval system*. CACM 26(11), p. 1022-1036, 1983.
[Sal&Buc., 1988] G. Salton and C. Buckley. Term-Weighting Approaches in Automatic Text Retrieval. Information Processing and Management. 1998. (5): pp 513-523.
[Saracevic., 1970] Saracevic, T. « The concept of “relevance” in information science :ahistoricalreview », dans T Saracevic (dir), Introduction to Information science. R.R.Bowker,New York, 1970.

[Saracevic., 1996] Saracevic, T. « Relevance reconsidered ».Conception of Library and InformationScience, 1996.

[Seb., 2002] Sebastiani, F.: Machine learning in automated text categorization. *ACM Computing Surveys*, 34(1),1–47, 2002.

[Sch&al ., 1995] H. Schütze, D. Hull and J. O. Pedersen. (1995). A Comparison of Classifiers and Document Representations for the Routing Problem. Proceedings of the 18th ACM International Conference on Research and Development in Information Retrieval (SIGIR 1995), E.A. FOX, P. INGWERSEN and R. FIDEL (edts.). ACM Press. July 1995, Washington, USA. pp 229-237.

[Slimani et al., 2006] Slimani T. Yaghlane B. B., and Mellouli K. A new similarity measure based on edge counting. In Proceedings of world academy of science, engineering and technology, Vol. 17 (December 2006).

[Sin., 1997] A. Singhal. Term Weighting Revisited. PhD thesis. Department of Computer Science, Cornell University. 1997.

[Singhal., 1997] A. K. Singhal, Term weighting revisited, PHD of Cornell University 1997.

[Sowa., 1984] J. Sowa. Conceptual Structures: Information Processing in Mind and Machine.

Addison Wesley, 1984. Ellis Horwood, 1992.

[Sparck Jones., 1979] K. Sparck Jones Experiments in relevant weighting of search terms.IPM 1979.

[Tambellini., 2007] C. Tambellini. "Un système de recherche d'information adapté aux données incertaines: adaptation du modèle de langue". Thèse de doctorat en informatique, Université de Nice-Sophia Antipolis-UFR sciences, 2007.

[Uschold&King., 1995] Towards a methodology for building ontologies, in Proceedings of the Workshop on Basic Ontological Issues in Knowledge Sharing, IJCAI'95, 1995.

[Van Heijst & Schreiber & Wielinga., 1997] Using explicit ontologies in kbs development. International Journal of Human and Computer Studies - Knowledge Acquisition, 46(2/3), 183–292.

[Wong et al., 1985] S. Wong, W. Ziarko, P.Wong. Generalized vector spaces model in information retrieval. In Proc. of the 8th ACM-SIGIR conference, p. 18-25. Montreal, Quebec, 1985.

[Wup., 1994] Z. Wu et M. Palmer. Verb semantics and lexical selection. In Proceedings of the 32nd Annual Meeting of the Associations for Computational Linguistics, pp 133- 138. 1994.

[Zargayouna et al., 2004] Zargayouna, H., Salotti, S. Mesure de similarité dans une ontologie pour l'indexation sémantique de documents XML. Actes de la conférence IC'2004.

[Zemirli., 2008] Wahiba Nesrine Zemirli, Modèle d'accès personnalisé à l'information basé sur les Diagrammes d'Influence intégrant un profil utilisateur évolutif . Thèse de Doctorat en Informatique de l'université de Toulouse, 2008.