

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

Mémoire de fin d'études

Pour l'obtention du diplôme de Master en Informatique

*Option: Intelligence Artificielle (M.I.D)*

*Thème*

**Enrichissement de la représentation  
conceptuelle dans la catégorisation du  
texte en utilisant les mesures de  
similarité sémantique**

**Réalisé par :**

- BOUYACOUB soumia
- KAOUADJI Sarra

*Présenté le 11 septembre 2013 devant le jury composé de MM.*

- Mr BENAMMAR Abdelkarim (Président)
- Mr BENTAALLAH Mohammed Amine (Encadreur)
- Mr SMAHI Mohammed Ismail (Examineur)
- Mr MOUFFAK Benattou (Examineur)

## Remerciement

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

Ce travail aussi modeste, n'a été rendu possible que grâce aux orientations éclairées de notre encadreur : *Monsieur A. BENTAALLAH* que nous tenons à lui exprimer notre parfaite gratitude et nos sincères remerciements pour la qualité de son encadrement et pour ses conseils judicieux et avisés.

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

Nous dédions ce mémoire à tous ceux qui ont contribué à notre formation et qui nous ont soutenus dans nos études

## Dédicace

Je dédie ce modeste mémoire à mes chers parents qui m'ont soutenu et encouragé tout le long de mes études,

À mon frère Salim, et mes sœurs, Lamia, Meryem, Fatima et Wissem.

À mon fiançais Sofiane et je le remercie pour tout.

Vous vous êtes dépensés pour moi sans compter.

En reconnaissance de tous les sacrifices consentis par tous et chacun pour me permettre d'atteindre cette étape de ma vie.

Avec toute ma tendresse.

À mes grands-parents, mes oncles et mes tantes

À mes cousins et cousines.

Vous avez de près ou de loin contribué.

Affectueuse reconnaissance.

À mes beaux-parents, à mon beau-frère et sa femme, à mes belles sœurs

À mon amie et binôme Soumia ainsi qu'à toute sa famille

À toute ma promo, et à mes amies, Amira, Hanane, Meryeme, Hayet, Sarah, Assia, Aicha, Imene, Amel, Tema et à leurs familles.

À Bambrik Ilyes,

Pour son aide. Sincère gratitude.

Je ne manquerai pas de remercier tous les professeurs qui m'ont suivi pendant le cursus universitaire.

Sarra, KAOUADJI

# **Dédicaces**

A mes très chers parents qui m'ont toujours soutenu et qui ont sacrifié leur vie afin de faire de moi ce que je suis. Que Dieu les protège.

A mon très cher mari, pour son soutien moral, sa gentillesse sans égal, son profond attachement m'ont permis de réussir mes études.

A mes sœurs.

A ma famille.

A mes amies.

A tous ceux qui nous sont chères et que nous aimons.

**BOUYACOUB Soumia**

# Table des matières

Liste des figures .....	4
Liste des tableaux.....	5
Introduction générale .....	6
Chapitre 1 : classification et catégorisation des textes.....	7
I. Introduction.....	8
II. Définition de la classification et la catégorisation des textes.....	8
III. Historique.....	9
IV. Le processus de la catégorisation des textes .....	9
IV.1 La représentation des textes .....	9
a. Représentation en sac de mots (bag of words).....	10
a. Représentation avec les racines lexicales.....	10
b. Représentation avec les lemmes.....	10
c. Représentation avec les n-gramme.....	11
b. Représentation conceptuelle.....	11
IV.2 La pondération des termes .....	11
a. Mesure TF (Term Frequency).....	11
b. Mesure TFIDF (Term Frequency Inverse Document Frequency) .....	11
IV.3 La réduction de la taille du vocabulaire .....	12
IV.4 Choix de classificateur .....	12
a. Machine à support vectoriel (SVM).....	12
b. K plus proches voisins .....	13
c. Méthode de Rocchio .....	13
d. Naïve bayes .....	14
a. Les arbres de décision .....	14
b. Les réseaux de neurone .....	14
IV.5 Evaluation du processus de catégorisation.....	15
V. Les applications de la catégorisation des textes.....	15
VI. Conclusion .....	16
Chapitre 2 : Etat d'art : les mesures de similarité .....	17
I. Introduction.....	18
II. Définition d'une ontologie .....	18
III. Les composants d'une ontologie.....	18
III.1 Les concepts.....	19

III.2	Les relations .....	19
III.3	Les fonctions .....	19
III.4	Les axiomes.....	19
III.5	Les instances .....	19
IV.	Définition de la mesure de similarité .....	20
V.	Notion de la distance sémantique.....	20
VI.	Les différents approches de mesure de similarité .....	20
VI.1	Approches basées sur les arcs .....	21
a.	Mesure de Wu & Palmer.....	21
b.	Mesure de Rada et al.....	22
a.	Mesure de Ehrig et al.....	22
b.	La mesure de Hirst-St.Onge.....	22
c.	La mesure de Zargayouna .....	23
VI.2	Approches basées sur les nœuds .....	23
a.	Mesure de Resnik.....	23
b.	Mesure de Lin .....	24
c.	Mesure de Seco et al.....	24
VI.3	Approches hybrides.....	24
a.	Mesure de Jiang et Conrath.....	24
b.	Mesure de Leacock et Chodorow.....	25
c.	Mesure de Li et al.....	25
d.	Mesure de FaITH .....	25
VI.4	Approches basées sur l'espace vectoriel .....	26
a.	Similarité de Cosine .....	26
b.	Similarité de Jaccard .....	26
c.	Similarité de Dice.....	26
VI.5	Mesure de similarité intentionnelle.....	27
a.	Mesure de Tversky .....	27
VII.	Domaine d'application .....	27
VII.1	Traitement du langage naturel (NLP).....	27
VII.2	Bioinformatique .....	28
VII.3	Web Services.....	28
VIII.	Conclusion .....	28
Chapitre 3 : Evaluation des mesures de similarité sémantique .....		29
I.	Introduction.....	30
II.	Architecture de notre travail.....	30

III.	Description des approches implémentées .....	31
III.1	Représentation en sac de mots .....	33
III.2	Transformation des mots en synsets .....	33
III.3	Représentation conceptuelle.....	35
III.4.	Enrichissement .....	35
III.5.	Classification avec Kppv.....	37
IV.	Technologies et outils de développement .....	38
IV.1.	Langage JAVA.....	38
IV.2.	Environnement de développement .....	39
IV.3.	WordNet.....	39
IV.4.	JWNL.....	40
IV.5.	Corpus utilisé .....	40
V.	Evaluation de notre travail .....	41
VI.	Discussion .....	43
VII.	Conclusion .....	44
	Conclusion générale.....	44

## **Liste des figures**

Figure 2.1 :Taxonomie des approches de mesure de similarité.....	18
Figure 2.2 : Similarité entre concepts selon Tversky.....	25
Figure 3.1: Processus de la première approche.....	32
Figure3.2 : Processus de la deuxième approche.....	33
Figure 3.3 : Combinatoire des sens.....	34
Figure 3.4 : Représentation matricielle d'un corpus.....	36



## **Liste des tableaux**

Tableau 3.1 : Caractéristiques du nombre de mots et de concepts dans WordNet.....	40
Tableau 3.2 : Caractéristiques du corpus utilisé.....	41
Tableau 3.3 : Précision et rappel pour la première approche.....	42
Tableau 3.4 : Précision et rappel pour la mesure de Wu Palmer.....	43
Tableau 3.5 : Précision et rappel pour la mesure de Lin.....	43

## **Introduction générale**

La révolution de l'internet a fait exploser les informations textuelles, qui sont un patrimoine vivant des entreprises, des administrations et des particuliers, il est devenu indispensable aux utilisateurs du web de trouver les documents pertinents, pour cette raison il devient de plus en plus important de disposer de solutions efficaces pour conserver, chercher et classer ces informations, afin d'assister les utilisateurs à trouver leurs besoins et faciliter leur travail dans certaines tâches qui sont devenues impossible à traiter manuellement. Donc il est très intéressant de compter sur une application automatique qui est la classification et la catégorisation des textes.

Cette thèse traite l'évaluation de l'utilisation des mesures de similarité sémantique pour la classification des textes, qui consiste à représenter les documents classés et non classés par une bonne méthode de représentation. L'objectif principal est de calculer la mesure de similarité entre les documents classés et le document non classé.

Nous avons décomposé notre mémoire en trois chapitres. Le premier chapitre vise à présenter le processus de la catégorisation des textes et les principales phases de ce dernier, ainsi, les applications liées à la catégorisation des textes, le deuxième chapitre présente un état d'art sur les mesures de similarité sémantique et leurs approches. Enfin le dernier chapitre qui expose la description des approches implémentées ainsi que les résultats obtenus.

# **Chapitre1 : classification et catégorisation des textes**

---

## I. Introduction

De nos jours, la classification des textes est un domaine de recherche très actif, permettant à l'utilisateur de trouver les informations pertinentes dans un temps raisonnable, afin de résoudre les problèmes d'accès à l'information voulu.

Dans ce chapitre, nous présentons d'abord un bref historique sur la catégorisation et la classification des textes, leur définition ainsi que leur processus. Nous exposons, ensuite les différents problèmes liés à la catégorisation des textes et nous présentons, à la fin, quelques applications utilisées dans ce domaine.

## II. Définition de la classification et la catégorisation des textes

Jalam définit dans [1] la catégorisation de textes comme étant la recherche d'une relation bijective qui consiste à "chercher une liaison fonctionnelle entre un ensemble de textes et un ensemble de catégories (étiquettes, classes)". C'est à dire associer une catégorie à un texte libre, en fonction des informations qu'il contient.

Sebastiani définit formellement dans [2] la catégorisation des textes comme le processus qui consiste à associer une valeur booléenne à chaque paire  $(d_j, c_i) \in D \times C$ , où  $D$  l'ensemble des textes et  $C$  l'ensemble des catégories. La valeur  $V$  (Vrai) est alors associée au couple  $(d_j, c_i)$  si le texte  $d_j$  appartient à la classe  $c_i$  tandis que la valeur  $F$  (Faux) lui sera associée dans le cas contraire. Le but de la catégorisation de texte est de construire une procédure (modèle, Classificateur) notée  $\Phi: D \times C \rightarrow \{V, F\}$  qui associe une ou plusieurs étiquettes (catégories) à un document  $d_j$  telle que la décision donnée par cette procédure coïncide le plus possible avec la fonction  $\Phi^{\wedge}: D \times C \rightarrow \{V, F\}$ , la vraie fonction qui retourne pour chaque vecteur  $d_j$  une valeur  $c_i$ .

## III. Historique

L'idée d'effectuer la classification des textes remonte au début des années 60 et qui a connu des progrès considérables à partir des années 90 avec l'apparition des algorithmes beaucoup plus performants qu'avant. Ces évolutions technologiques et algorithmes avancées font aujourd'hui de la catégorisation, un outil fiable.

Les termes ‘classification’ et ‘catégorisation’ ont des histoires et des origines très différentes. Le terme classification est apparu pour la première fois dans la cinquième édition du dictionnaire de l’Académie Française en 1798 [3] sous la définition : « distribution en classes et suivant un certain ordre » et dans la dernière édition [4] par « l’Action de classer et le résultat de cette action ». Le terme ‘catégorisation’ n’existe pas dans le dictionnaire de l’Académie Française, contrairement au mot ‘catégorie’ qui est défini dans tous les éditions du dictionnaire comme étant une classe dans laquelle on range plusieurs choses qui sont des espèces différentes, mais qui appartiennent à un même genre [4].

#### **IV. Le processus de la catégorisation des textes**

D’une façon générale, le processus de catégorisation reçoit en entrée un texte afin de lui associer sa catégorie en sortie. Pour identifier la catégorie d’un texte, un ensemble d’étapes est habituellement suivies. D’après Jalam dans [1], ces étapes sont :

- La représentation des textes
- La pondération des termes
- La réduction de la taille de vocabulaire
- Choix de classificateur
- Evaluation du modèle

##### **IV.1 La représentation des textes**

La représentation des textes est la phase la plus importante dans le processus de catégorisation des textes, et cela pour classer les textes d’une manière efficace. Cette représentation consiste à représenter chaque document sous forme d’un vecteur, dont les composants sont les termes de ce document. Les différentes méthodes pour la représentation des textes sont :

###### **a. Représentation en sac de mots (bag of words)**

Cette méthode consiste à représenter chaque document par un vecteur, dont les composants sont les mots contenus dans le texte. L’analyse lexicale est un processus qui permet de convertir le texte d’un document en un ensemble de mots qui permet de reconnaître les espaces de séparation des mots, les signes de ponctuation, ... etc. ; ces derniers sont supprimés de la représentation.

L'avantage de cette représentation est d'exclure toute analyse grammaticale et toute notion de distance entre les mots, mais l'inconvénient est que le regroupement des mots d'un document sans prendre en compte les combinaisons et l'ordre des mots dans la phrase entraîne une perte dans la sémantique du texte.

#### **b. Représentation avec les racines lexicales**

Cette représentation consiste à regrouper les mots de la même racine dans une seule composante, alors que cette racine peut être commune pour des mots qui ont des sens différents. On a par exemple les mots «descendre» «descendance » et «descendu» ont la même racine «descend » mais leurs notions sont différentes. Cette méthode se fait grâce à des algorithmes qui ont été proposés; l'un des plus connus pour la langue anglaise est l'algorithme de Porter [5].

#### **c. Représentation avec les lemmes**

Cette représentation est compliquée à mettre en œuvre puisqu'elle nécessite une analyse grammaticale des textes afin de remplacer tous les verbes du document par leur forme infinitive et les noms par leur forme au singulier. L'objectif de cette représentation est d'associer à chaque mot, une entrée dans le lexique qui est définie comme un ensemble de lemmes. Par exemple le lemme de « chantaient » est « chanter ». Cette méthode est simple mais elle peut causer plusieurs problèmes comme la perte de l'information donnée par le contexte syntagmatique, nécessaire à la distinction des lemmes polysémiques et la présence de synonymes, considérés comme des lemmes différents même s'ils font référence au même concept [6].

#### **d. Représentation avec les n-gramme**

Cette représentation consiste à représenter le document par des n-grammes qui sont une séquence de n caractères consécutifs. Plusieurs travaux ont montré l'efficacité des n-grammes comme méthode de représentation des textes pour la classification. Cette méthode a plusieurs avantages, comparativement à d'autres techniques, Les n-grammes capturent automatiquement les racines des mots les plus fréquents sans passer par l'étape de recherche des racines lexicales [7]. Il introduit aussi la notion d'indépendance de la langue comme montré dans [8]. Les espaces sont pris en considération parce qu'en effet, la non prise en compte de ces derniers introduit du bruit.

### e. Représentation conceptuelle

Cette étape sert à représenter le document textuel sous forme d'un ensemble de concepts qui est un groupe de synonymes appelés «*synsets*». Selon Rehel dans [9], l'avantage de la représentation conceptuelle est de réduire l'espace de travail car les mots qui sont synonymes partagent au moins un concept. L'inconvénient de cette représentation est qu'il n'existe pas des bases lexicales pour toutes les langues.

## IV.2 La pondération des termes

La pondération des termes permet de mesurer l'importance d'un terme dans un document. L'objectif est de trouver les termes qui représentent le mieux le contenu d'un document. Les méthodes les plus populaires sont :

### a. Mesure TF (Term Frequency)

Ce facteur permet de mesurer la représentativité locale d'un terme. Il prend en compte les informations locales du terme qui ne dépendent que du document donné, et indique l'importance du terme dans ce document.

### b. Mesure TFIDF (Term Frequency Inverse Document Frequency)

La mesure *Tf-Idf* est une bonne approximation de l'importance d'un terme dans un document, particulièrement dans des corpus de documents de tailles homogènes. Le poids d'un terme  $T$  dans un document  $D$  est calculé comme suit :

$$\mathbf{TFIDF(T, D)} = \mathbf{TF(T, D)} * \log (\mathbf{N} / \mathbf{DF(T)})$$

Avec :

**TF(T, D)** : la fréquence du terme dans le document,

**N** : le nombre total de documents de la base documentaire et

**DF(T)** : le nombre de documents contenant le terme.

## IV.3 La réduction de la taille du vocabulaire

Un des problèmes de la catégorisation de textes est l'espace mémoire nécessaire pour traiter des corpus de grandes tailles, donc il est difficile de prendre l'ensemble de tous les mots comme étant des attributs, car cela engendre une perte de mémoire et de temps de calcul. Pour cela il existe une approche qui est la réduction des dimensions permettant de résoudre ce problème. Son objectif est de

sélectionner ou d'extraire un sous-ensemble optimal de caractéristiques pertinentes pour un critère fixé. Les techniques de réduction de la taille du vocabulaire sont classées en deux catégories :

- *Sélection d'attributs* : (*feature selection*)

Une des méthodes la plus simple et la plus évidente consiste à prendre les attributs d'origine, conserver seulement ceux utiles à la catégorisation selon une fonction d'évaluation et éliminer des éléments du vocabulaire qui ne nous intéressent pas.

- *Extraction d'attributs* : (*feature extraction*)

À partir des attributs de départ, elles créent de nouveaux attributs, en faisant soit des regroupements ou des transformations.

#### IV.4 Choix de classificateur

Dans le domaine de la catégorisation de textes, différents types de classificateurs ont été mis au point. Nous présentons ci-dessous quelques algorithmes d'apprentissage couramment utilisés dans ce cadre :

##### a. Machine à support vectoriel (SVM)

Ces machines, proposées par Vapnik dans [10] ont été utilisées avec succès dans plusieurs tâches d'apprentissage. Le but de SVM est de trouver un classificateur qui sépare au mieux les données et maximise la distance entre ces deux classes. Avec SVM, ce classificateur est un classificateur linéaire appelé hyperplan. Dans le cas de la catégorisation des textes, les entrées sont des documents et les sorties sont des catégories. SVM est considéré parmi les plus performants pour la catégorisation en raison de sa modélisation simpliste et rapide à calculer par une machine étant donné que SVM est un classificateur linéaire. Par contre elle introduit des concepts complexes peu adaptés aux corpus de grandes tailles non fixes, sans oublier de rappeler de son faible pouvoir descriptif puisque les coefficients ne sont pas interprétables intuitivement par des humains.

##### b. K plus proches voisins

K-plus proches voisins « K-Nearest Neighbour » est une méthode qui a prouvé son efficacité face au domaine de la catégorisation des textes. Ses performances la



situent parmi les meilleures méthodes de catégorisation. L'idée de base de l'algorithme des k-plus proches voisins (kPPV), est de représenter chaque texte dans un espace vectoriel, dont chacun des axes représente un élément textuel. Les K éléments les plus proches sont sélectionnés et le document est assigné à la classe majoritaire.

Les k-PPV ont été utilisés en CT pour la première fois par B. Masand. Y. Yang qui a montré que l'algorithme est parmi les meilleurs actuellement [11]. La distance entre un texte et ses voisins se fait par une métrique qui est la distance. Cette métrique peut être calculée comme suit :

- **Mesure Cosinus**

Le cosinus est l'une des premières mesures à avoir été utilisée dans le domaine de la recherche d'information. Le cosinus entre deux vecteurs est obtenu en calculant le produit scalaire entre ces deux vecteurs, que nous divisons par le produit de la norme des deux vecteurs. Le cosinus entre deux vecteurs a et b est défini par l'équation suivante :

$$\text{Cosinus (a, b)} = \frac{\sum (a \times b)}{\sqrt{\sum a^2 \times \sum b^2}}$$

**c. Méthode de Rocchio**

Le classificateur de Rocchio paru dans [12], est l'un des plus simples et plus anciens algorithmes de classification du modèle vectoriel. Ce classificateur a été largement utilisé dans la catégorisation textuelle. L'avantage de ce type de classificateur est la simplicité et l'interprétabilité. L'apprentissage de ce type de classificateur est souvent précédé par une sélection et une réduction de termes.

**d. Naïve bayes**

L'algorithme Naïve Bayes (NB), est une méthode très connue en apprentissage, elle permet de calculer les probabilités conditionnelles. Dans le cas de la classification de textes, on cherche la classification qui maximise la probabilité d'observer les mots du document. Lors de la phase d'entraînement, le classificateur calcule les probabilités qu'un nouveau document appartienne à telle catégorie à partir de la proportion des documents d'entraînement appartenant à cette catégorie. Il calcule aussi la probabilité

qu'un mot donné soit présent dans un texte, sachant que ce texte appartient à telle catégorie.

Quand un nouveau document doit être classé, on calcule les probabilités qu'il appartienne à chacune des catégories à l'aide de la règle de Bayes. [9].

#### e. Les arbres de décision

Les arbres de décision sont composés d'une structure hiérarchique en forme d'arbre. Un arbre de décision est un graphe orienté, sans cycles, dont les nœuds portent une question, les arcs des réponses, et les feuilles des conclusions. Un classificateur de texte basé sur la méthode d'arbre de décision est un arbre de nœuds internes qui sont marqués par des termes, les branches qui sortent des nœuds sont des tests sur les termes, et les feuilles sont marquées par catégories [13]. Une méthode pour effectuer l'apprentissage d'un arbre de décision pour une catégorie  $C_i$  consiste à vérifier si tous les exemples d'apprentissage ont la même étiquette. Dans le cas contraire, nous sélectionnons un terme  $T_k$ , et nous partitionnons l'ensemble d'apprentissage en classes de documents qui ont la même valeur pour  $T_k$ , et à la fin nous créons les sous-arbres pour chacune de ces classes. Ce processus est répété récursivement sur les sous-arbres jusqu'à ce que chaque feuille de l'arbre généré de cette façon contienne des exemples d'apprentissage attribués à la même catégorie  $C_i$ , qui est alors choisie comme l'étiquette de la feuille.

#### f. Les réseaux de neurone

Les approches neuronales furent les premières à être utilisées afin de réaliser un apprentissage de type statistique grâce à leur capacité de classification et de généralisation. Un réseau de neurones artificiels est composé d'une ou de plusieurs couches se succédant dont chaque entrée est la sortie de la couche qui la précède. Chaque couche  $i$  est composée de  $N_i$  neurones, prenant leurs entrées sur les  $N_{i-1}$  neurones de la couche précédente. Nous pouvons ainsi voir un réseau de neurones artificiels comme un réseau ou graphe orienté dont les nœuds sont les neurones artificiels. Le but va être d'attribuer des poids synaptiques à chaque neurone afin d'obtenir le résultat voulu en sortie. [14].

#### IV.5 Evaluation du processus de catégorisation

Afin d'apprécier tout processus de catégorisation, il est important d'utiliser une méthode d'évaluation. Cette dernière doit comporter la précision et le rappel. La précision est défini comme la probabilité conditionnelle ; le rappel mesure la largeur de l'apprentissage et correspond à la fraction des documents pertinents, parmi ceux fournis par le classificateur.

**Précision** =  $A / (A + B)$  pour  $A + B > 0$

**Rappel** =  $A / (A + C)$  pour  $A + C > 0$

Avec :

A : le nombre de documents correctement attribués à la catégorie.

B : le nombre de documents incorrectement attribués à la catégorie.

C : le nombre de documents qui auraient dû lui être attribués mais qui ne l'ont pas été.

#### V. Les applications de la catégorisation des textes

La catégorisation des textes est utilisée dans de nombreux domaines, apportant de nombreux avantages tel que :

- Reconnaissance :
  - identification de la langue du document,
  - Filtrage et détection des spams.
- Gestion rapide et efficace du flux d'information afin de satisfaire le client (recueil et acheminement des courriers vers leurs destinataires.
- désambiguïsation des termes.
- catégorisation des documents multimédia.
- l'indexation automatique des textes.

#### VI. Conclusion

Le processus de catégorisation est un domaine très actif apportant à l'utilisateur de nombreuses applications avantageuses et intéressantes.

Dans ce chapitre, nous avons présenté le processus de catégorisation des textes avec ses différentes approches ainsi que les applications de la catégorisation des textes.

Dans le chapitre suivant, nous allons présenter un état d'art sur les mesures de similarité ainsi que son domaine d'application.

# **Chapitre2 : les mesures de similarité sémantique**

---

## I. Introduction

Le domaine de l'identification de la similarité est un sujet fondamental qui est adopté par plusieurs techniques telles que le Web sémantique et en particulier, le domaine de la recherche de l'information. Ce dernier repose largement sur des mesures pour l'identification de la similarité entre les documents [15] [16].

A travers ce chapitre, nous présentons d'abord une définition de l'ontologie ainsi que ses composantes, ensuite une définition de la mesure de similarité et de la distance sémantique et en dernier, nous citons les approches de mesures de similarité ainsi que leur domaine d'application.

## II. Définition d'une ontologie

Le concept ontologie existe depuis très longtemps, notamment en philosophie. Il vient du grec : *onto* qui est l'étude de l'être en tant qu'être et *logo* qui signifie l'univers, c'est-à-dire l'étude des propriétés générales de ce qui existe.

D'une manière générale une ontologie constitue une approche très efficace pour représenter les connaissances. D'après Gruber [17] une ontologie est une spécification explicite de la conceptualisation.

## III. Les composants d'une ontologie

Une ontologie est formée par des concepts et des relations entre ceux-ci. Ainsi, une ontologie peut être vue comme un treillis de concepts et de relations entre ces concepts destinés à représenter les objets du monde sous une forme compréhensible aussi bien par les hommes que par les machines.

Si certaines divergences relatives à la structure (degré de la formalisation) de l'ontologie ont été constatées, les composants d'une ontologie sont toujours les mêmes : une ontologie est constituée des concepts et des relations ainsi que des propriétés et des axiomes.

### III.1 Les concepts

Ce sont des notions (ou objets) permettant la description d'une tâche, d'une fonction d'une action, d'une stratégie ou d'un processus de raisonnement, etc. Ils peuvent être abstraits ou concrets, élémentaires ou composés, réels ou fictifs. Habituellement, les concepts sont organisés en taxonomie. Une taxonomie

est une hiérarchie de concepts (ou d'objets) reliés entre eux en fonction de critères sémantiques particuliers.

### III.2 Les relations

Ce sont les liens organisant les concepts de façon à représenter un type d'interaction entre les concepts d'un domaine, elles sont formellement définies comme tout sous-ensemble d'un produit de  $n$  ensembles, c'est-à-dire

$R: C_1 \times C_2 \times \dots \times C_n$ . Des exemples de relations binaires sont : *sous-concept-de*, *connecté-à*, *sorte-de*, etc.

### III.3 Les fonctions

Ce sont des cas particuliers de relations dans lesquelles le  $n^{\text{ième}}$  élément de la relation est unique pour les  $n-1$  précédents.

Formellement, les fonctions sont définies ainsi,  $F: c_1 * c_2 * \dots * c_{n-1} c_n \longrightarrow$

Comme exemple de fonction binaire, nous avons la fonction *mère de*.

### III.4 Les axiomes

Les axiomes de l'ontologie permettent de définir la sémantique des termes (classes, relations), leurs propriétés et toutes contraintes quant à leur interprétation. Ils sont définis à l'aide de formules bien formées de la logique du premier ordre en utilisant les prédicats de l'ontologie.

### III.5 Les instances

Elles constituent la définition extensionnelle de l'ontologie ; elles sont utilisées pour représenter des éléments dans un domaine. *Exemple*: les individus *Peugeot*

*206* et *Atos* sont des instances du concept « Voiture ».

## IV. Définition de la mesure de similarité sémantique

L'objectif des mesures de similarité sémantique est d'évaluer la proximité sémantique entre les concepts. Le calcul de similarité entre deux concepts permet de déterminer s'ils sont similaires c'est-à-dire s'ils atteignent un certain niveau de ressemblance ou dissimilaire qui peuvent être également liés sémantiquement par des relations lexicales : antonymie, spécialisation, etc.

## V. Notion de la distance sémantique

Une distance  $d$  est sémantique si et seulement si elle vérifie les propriétés suivantes :

- Si  $A \equiv B$  alors  $d(A, B) = 0$  : la distance entre deux concepts équivalents est nulle.
- Si  $A \sqsubseteq B \sqsubseteq C$  alors  $d(A, B) \leq d(A, C)$  : la distance entre un concept avec un subsumant direct est inférieure à sa distance avec n'importe quel autre subsumant.
- Si  $A \cap B$  alors  $d(A, B) = \infty$  : la distance entre deux concepts incompatibles est infinie.

Cette distance  $d$  est sémantique car elle permet d'exprimer certaines relations sémantiques intuitives entre les concepts, relative à la ressemblance entre les concepts, comme le fait que la distance entre deux concepts équivalents est nulle.

## VI. Les différents approches de mesure de similarité

Dans cette partie nous définissons les approches principales de la mesure de similarité. Ainsi la figure montre une partie de classification et quelques approches d'échantillon.

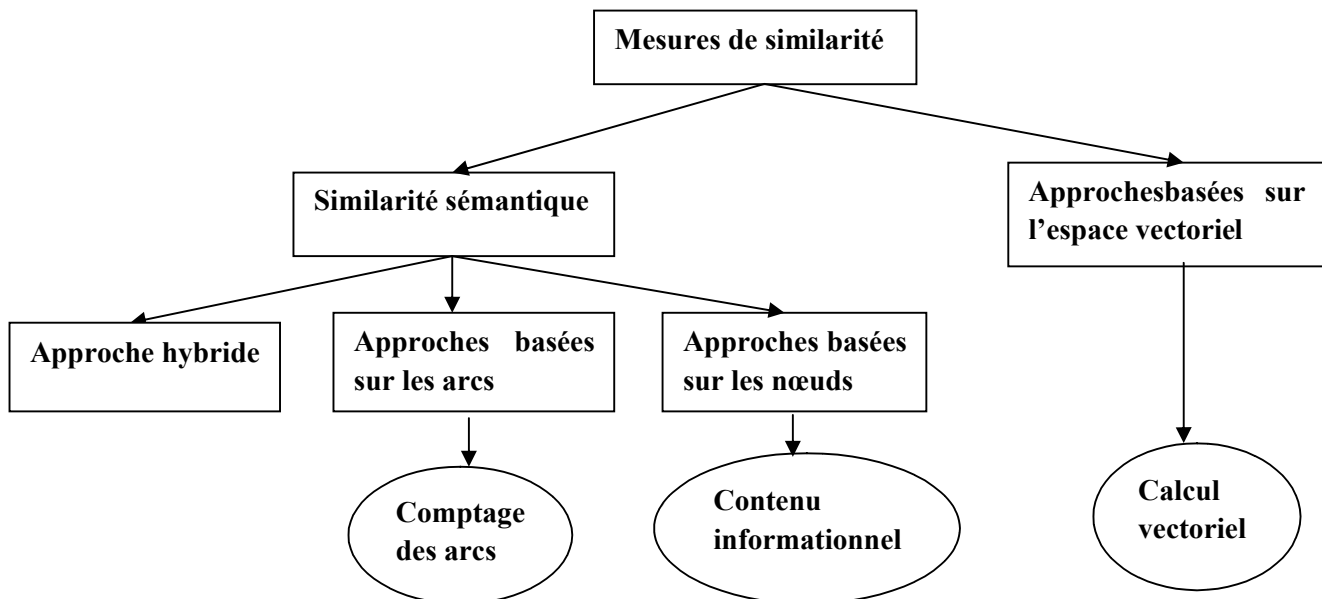


Figure 2.1 : Taxonomie des approches de mesure de similarité.

### VI.1 Approches basées sur les arcs

Ce type de mesure se sert de la structure hiérarchique de l'ontologie qui est représentée par un graphe dont les nœuds sont des concepts, et les arcs sont



les liens entre ces concepts, et cela pour déterminer la similarité sémantique entre les concepts qui peut être calculée à partir du nombre de liens qui séparent les deux concepts. Parmi les travaux classifiés sous cette approche on peut citer :

### a. Mesure de Wu & Palmer

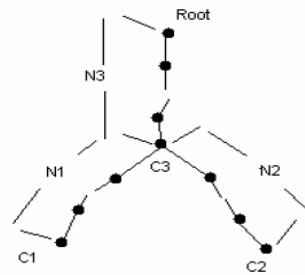
Dans une ontologie, la similarité est définie par rapport à la distance qui sépare deux concepts dans la hiérarchie et également par leur position par rapport à la racine.

La similarité de Wu & Palmer [18] entre  $c_1$  et  $c_2$  est :

$$\text{sim}_{\text{WPalmer}}(c_1, c_2) = \frac{2 * \text{prof}(c)}{\text{dist}(c_1, c) + \text{dist}(c_2, c) + 2 * \text{prof}(c)}$$

Où  $c$  est le concept le plus spécifique qui subsume les deux concepts  $c_1$  et  $c_2$  c'est-à-dire le PPG qui sépare  $c_1$  et  $c_2$ ,  $\text{prof}(c)$  est le nombre d'arcs qui sépare  $c$  de la racine et  $\text{dist}(c_i, c)$  le nombre d'arcs qui séparent  $c_i$  de  $c$ .

$$\text{ConSim}(C1, C2) = \frac{2 * N3}{N1 + N2 + 2 * N3}$$



Cette mesure a l'avantage d'être simple à implémenter et d'avoir d'aussi bonnes performances que les autres mesures de similarité [19].

### b. Mesure de Rada et al.

Cette mesure de similarité [20] dans un réseau sémantique peut être calculée en se basant sur les liens hiérarchiques «is-a». Pour calculer la similarité entre deux concepts dans une ontologie, on doit calculer la distance entre les concepts c'est-à-dire le nombre des arcs minimums qui les séparent par le chemin le plus court. C'est un moyen pour évaluer la similarité sémantique dans une ontologie hiérarchique.

$$\text{sim}_{\text{rada}}(c_1, c_2) = \frac{1}{1 + \text{dist}(c_1, c_2)}$$

La mesure de [20] utilise une métrique  $dist(c_1; c_2)$ , qui indique le nombre d'arcs minimum à parcourir pour aller d'un concept  $c_1$  à un concept  $c_2$ .

### c. Mesure de Ehrig et al.

Ehrig introduit dans [21] un travail de mesure de similarité pour les ontologies. Ce travail introduit trois couches : les données, l'ontologie et le contexte. La similarité des entités est mesurée au niveau des données en considérant les valeurs de données de type simple ou complexe (entiers, caractères). Les relations sémantiques entre les entités sont mesurées au niveau de la couche de l'ontologie. Finalement la couche du contexte spécifique comment les entités de l'ontologie sont utilisées dans un certain contexte externe, plus spécifiquement, le contexte de l'application.

### d. La mesure de Hirst-St. Onge

La mesure de Hirst-St. Onge dans [22] prend en considération toutes les relations dans WordNet. Les liens sont classés comme *haut* (eg. partie-de), *bas* (eg. sous-classe), *horizontal* (eg. antonyme). La similarité est calculée entre mots par le poids du chemin le plus court qui mène d'un terme à un autre. Il est calculé en fonction de ces classifications qui indiquent les changements de direction :

$$Sim(c_1, c_2) = T - \text{chemin} - K \times d$$

Tels que  $T$  et  $K$  sont des constantes, *chemin* est la longueur du chemin le plus court en nombre d'arcs et  $d$  est le nombre de changements de direction. L'idée est que deux concepts sont proches sémantiquement si leurs synsets sont connectés par un chemin qui n'est pas très long et qui ne change pas souvent de direction. S'il n'y a pas de chemin, le poids est égal à zéro.

### e. La mesure de Zargayouna

La mesure de [23] est inspirée de celle de [18]. Le lien père-fils est ainsi privilégié par rapport aux autres liens de voisinage en adaptant la mesure de Wu-Palmer. L'adaptation de la mesure est faite au travers de la fonction de calcul du degré de spécialisation d'un concept (*spec*) qui mesure sa distance par rapport à l'anti-racine.

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

$$\text{spec}(c_1, c_2) = \text{prof}_b(c) * \text{dist}(c_1, c) * \text{dist}(c_2, c)$$

Où  $\text{prof}_b(c)$  correspond au nombre maximum d'arcs qui séparent le plus petit ancêtre commun du concept « virtuel » représentant l'anti-racine.

## VI.2 Approches basées sur les nœuds

Cette approche prend en considération le contenu informatif (IC) des concepts de l'ontologie. La similarité est alors calculée à partir de l'information partagée par les concepts. Le contenu informatif est défini par :

$$\text{IC}(c) = -\log p(c).$$

Où nous avons  $c$  un concept, et  $p(c)$  la probabilité de lui trouver un de ses descendants dans le corpus. Parmi les mesures basées sur le contenu informationnel on peut citer :

### a. Mesure de Resnik

La notion du contenu informationnel (IC) a été initialement introduite par Resnik dans [24] qui a prouvé que la similarité sémantique entre deux concepts est mesurée par la quantité de l'information qu'ils partagent. Cette information partagée est égale au contenu informationnel du plus petit généralisant (PPG) c'est à dire le concept le plus spécifique qui subsume les deux concepts dans l'ontologie. La formule proposée par Resnik est définie par:

$$\text{Sim}(c_1, c_2) = \text{CI}(\text{ppg}(c_1, c_2))$$

Cette mesure est un peu sommaire car elle ne dépend que du concept le plus spécifique.

### b. Mesure de Lin

Lin a défini [25] dans une mesure de similarité légèrement différente de celle de Resnik :

$$\text{sim}_{\text{Lin}}(c_1, c_2) = \frac{2 * \text{CI}(\text{LCS}(c_1, c_2))}{\text{CI}(c_1) + \text{CI}(c_2)}$$

Cette mesure se base aussi sur le concept d'IC et elle utilise une approche hybride qui combine deux sources de connaissances différentes (Thesaurus, corpus). Les travaux de Mil [26] ont évalué cette mesure à travers une expérience qui utilise des sujets humains pour évaluer la similarité entre

30 paires de noms, il en ressort que cette méthode offre une amélioration significative.

### c. Mesure de Seco et al.

Cette méthode calcule le contenu informatif des nœuds à partir de WordNet au lieu d'un corpus. Seco et al. [27] utilisent les hyponymes descendants des concepts pour calculer le contenu informatif de ceux-ci.

$$\frac{\log\left(\frac{\text{hypo}(c)+1}{\text{max}_{wn}}\right)}{\log\left(\frac{1}{\text{max}_{wn}}\right)} = 1 - \frac{\log(\text{hypo}(c) + 1)}{\log(\text{max}_{wn})}$$

Où :  $\text{hypo}(c)$  est le nombre d'hyponymes du concept  $c$  ; et  $\text{max}_{wn}$  : une constante qui indique le nombre de concepts de la taxonomie.

## VI.3 Approches hybrides

Ces approches sont fondées sur un modèle mixte qui combine les approches basées sur les arcs et les approches basées sur les nœuds c'est-à-dire le contenu informationnel qui est considéré comme facteur de décision.

### a. Mesure de Jiang et Conrath

Jiang et Conrath [28] ont apporté une nouvelle formule qui consiste à combiner le contenu informationnel du PPG à ceux des concepts dont on cherche la similarité, elle prend en considération aussi le nombre d'arcs. La mesure adoptant cette méthode est basée sur la combinaison d'une source de connaissance riche (thesaurus) avec une source de connaissance pauvre (corpus). Notons que cette formule est définie par l'inverse de la distance sémantique.

$$\text{Sim}(c_1, c_2) = 1/\text{distance}(c_1, c_2)$$

Sachant que la distance entre  $c_1$  et  $c_2$  est calculée par la formule suivante :

$$\text{distance}(c_1, c_2) = \text{CI}(c_1) + \text{CI}(c_2) - (2 \cdot \text{CI}(\text{ppg}(c_1, c_2)))$$

### b. Mesure de Leacock et Chodorow

La mesure de Leacock et Chodorow [29] est une mesure basée sur la longueur du plus court chemin entre les synsets de WordNet. Le plus court chemin est celui qui comprend le plus petit nombre de nœuds intermédiaires.

Cette mesure est définie comme suit :

$$Sim_{litch}(c_1, c_2) = \max [-\log(\text{length}(c_1, c_2) / (2 \cdot D))] ]$$

Où  $\text{length}(c_1, c_2)$  est le plus court chemin entre deux nœuds et  $D$  la profondeur maximale dans la taxonomie (égale à 16 dans WordNet 1.7).

### c. Mesure de Li et al.

Li et al. Proposent en 2003 [30] une mesure qui combine à la fois la distance taxonomique ( $l = N_1 + N_2$ ), la profondeur du concept commun le plus spécifique dans la taxonomie ( $h = N_3$ ) ainsi que la densité sémantique locale ( $d = IC(lso(s_1, s_2))$ ), cette dernière étant exprimée en terme de contenu informationnel. Leur mesure est exprimée par :

$$Sim_{Li}(s_1, s_2) = f(f_1(l), f_2(h), f_3(d))$$

Où  $f_1$ ,  $f_2$  et  $f_3$  sont les fonctions de transfert non-linéaires respectives pour Chaque type d'information.

### d. Mesure de Faith

Cette mesure locale combine des aspects de différents types, est proposée par [31] sous la forme de l'extension des mesures à base de contenu informationnel.

$$Sim_{Faith} = \frac{IC(lso(s_1, s_2))}{IC(s_1) + IC(s_2) - IC(lso(s_1, s_2))}$$

Les traits communs aux concepts sont représentés par le contenu informationnel du concept commun le plus spécifique, et les traits spécifiques à un concept sont représentés par la différence entre le contenu informationnel de ce concept auquel on soustrait le contenu informationnel du concept commun le plus spécifique.

## VI.4 Approches basées sur l'espace vectoriel

Ces approches utilisent un vecteur caractéristique, dans un espace dimensionnel pour représenter chaque objet et calculent la similarité en se basant sur la mesure de cosinus ou la distance euclidienne. La définition de la similarité entre deux vecteurs d'objets est obtenue par leurs contenus internes. Parmi les approches citées dans la littérature on peut citer :

### a. Similarité de Cosine

Cette mesure utilise la représentation vectorielle complète, c'est-à-dire la fréquence des objets (mots). Deux objets (documents) sont similaires si leurs vecteurs sont confondus. Si deux objets ne sont pas similaires, leurs vecteurs forment un angle  $(X, Y)$  dont le cosinus représente la valeur de la similarité. La formule est définie par le rapport du produit scalaire des vecteurs  $x$  et  $y$  et le produit de la norme de  $x$  et de  $y$ .

$$\text{Simc}(X, Y) = \cos(X, Y) = \frac{x \cdot y}{\|x\|_2 \cdot \|y\|_2}$$

### b. Similarité de Jaccard

La mesure de similarité de Jaccard est définie par le nombre des objets communs divisé par le nombre total des objets moins le nombre d'objets communs :

$$\text{Simj} = \frac{x \cdot y}{\|x\|_2^2 + \|y\|_2^2 - x \cdot y}$$

Tels que  $x$  et  $y$  sont des vecteurs extraits à partir des concepts  $X$  et  $Y$ .

$\|x\| = \sqrt{\sum_{i=1}^n |x_i|^2}$  désigne la norme du vecteur  $x$  et  $\|x\|_2 = \sqrt{\sum_{i=1}^n |x_i|^2}$

### c. Similarité de Dice

La similarité de Dice est définie par le nombre des objets communs multipliés par 2 sur le nombre total d'objets. La mesure de Dice est donc définie par la formule suivante :

$$\text{Simd}(X, Y) = \frac{2 \cdot x \cdot y}{\|x\|_2^2 + \|y\|_2^2}$$

## VI.5 Mesure de similarité intentionnelle

D'un point de vue ensembliste, deux entités sont similaires si le cardinal de l'intersection des ensembles de leurs caractéristiques est plus grand que celui des sous-ensembles restants.

### a. Mesure de Tversky

Cette approche a été notamment développée par le psychologue A. Tversky qui a proposé dans [32] la définition suivante d'une similarité entre deux concepts :

$$\text{sim}_{\text{tversky}}(A, B) = \alpha \cdot \text{comm}(A, B) - \beta \cdot \text{diff}(A, B) - \gamma \cdot \text{diff}(B, A)$$

Ou  $\alpha, \beta, \gamma$  sont des constantes. Si  $\beta \neq 0$  et  $\alpha = 1$ , la similarité entre A et B correspond à la quantité de propriétés en commun. Si  $\alpha = 0, \beta > 0$  et  $\gamma > 0$  les entités A et B sont évaluées selon ce qui les différencie, nous avons alors une mesure de dissimilarité.



Figure 2.2: Similarité entre concepts selon Tversky[32]

Cette mesure est dépendante du cardinal des propriétés de chaque concept.

Une autre version de la formule précédente permet de palier ce problème :

$$sim_{tversky}(A, B) = \frac{comm(A, B)}{\beta \cdot diff(A, B) + \gamma \cdot diff(B, A) + comm(A, B)}$$

## VII. Domaine d'application

Afin de connaître l'importance de la mesure de similarité, il est nécessaire de rappeler les indications de son utilisation.

### VII.1 Traitement du langage naturel (NLP)

Dans cette indication on peut trouver : le travail de Pat dans [33] qui utilise la métrique de la similarité sémantique pour mesurer la similarité sémantique entre tous les sens d'un mot d'une paire donnée de mot et les désambigüiser ainsi dans un contexte donné. Les auteurs du travail [34] ont appliqué les mesures de similarité sémantique de WordNet pour évaluer la pertinence des expressions. Le travail de Hib [35] étudie l'utilité de la similarité sémantique dans le problème de correction d'orthographe, où des erreurs d'orthographe réelles sont détectées et corrigées automatiquement.

### VII.2 Bioinformatique

Le travail de Lor [36] s'intéresse à la similarité sémantique entre les protéines, plutôt que les termes de l'ontologie GO, c'est pour cette raison qu'il a combiné entre trois mesures de similarités [24][25] [28].

### VII.3 Web Services

Dans le travail de [37], il y a une proposition des métriques pour mesurer la similarité des services sémantiques annotés avec une ontologie OWL. La

mesure de similarité proposée est basée sur l'intuition que les objets similaires partagent les informations descriptives les plus communes.

### **VIII. Conclusion**

Dans ce travail nous avons classifié trois grandes familles d'approches pour l'identification de la similarité 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'approches sur les distances des arcs qui est basé sur le plus court chemin entre les nœuds. Finalement l'approche hybride qui combine entre les deux premières approches. Avec cette approche, il y a plusieurs manières de déterminer la similarité conceptuelle de deux mots dans un réseau sémantique hiérarchique.



# **Chapitre 3 : Evaluation des mesures de similarité**

---

## I. Introduction

Aujourd'hui La quantité d'information accessible sur le net est phénoménale, et sa catégorisation reste l'une des tâches les plus importantes. La catégorisation textuel utilise souvent le model vectoriel qui se base sur des mesures de similarités statiques. Le problème du modèle vectoriel c'est qu'il ne prend pas en considération la relation entre les composants du vecteur. Notre travail consiste à effectuer un enrichissement qui se base sur l'utilisation des mesures de similarités sémantiques.

Dans ce chapitre, nous présentons une vue générale et schématique des étapes suivies de notre approche ainsi que de leur description, afin de réaliser notre processus en se basant sur des stratégies de mesures de similarités sémantique, combinant en quelque sorte les idées des deux chapitres précédents.

## II. Architecture de notre travail

L'objectif de notre travail est de trouver une liaison fonctionnelle entre un ensemble de documents et un ensemble de catégories. Pour cela, il est nécessaire de disposer d'un ensemble de documents classé dans une catégorie à partir du quels nous faisons quelques étapes d'évaluations pour associer automatiquement une catégorie à n'importe quel document non classé. Dans le processus de la catégorisation des textes, nous nous sommes intéressées à l'étape de la représentation des textes En effet une bonne représentation donnera de bons résultats de classification.

De ce fait nous avons implémenté deux méthodes de représentation. La première utilise la représentation en "sac de mots" comme méthode de représentation, Les documents seront représentés par des vecteurs de mots. La deuxième est basée sur le WordNet pour traiter les documents classés et les documents non classés à fin de faire la représentation conceptuelle dans laquelle l'unité de vecteur serait un concept (groupe des synonymes). Ensuite nous passons à l'étape la plus importante de notre travail qui consiste à effectuer un enrichissement pour la représentation conceptuelle. Un enrichissement qui introduit le calcul des mesures de similarité sémantique sur des concepts qui n'appartiennent pas au dictionnaire des concepts d'apprentissage. En dernier, nous choisissons une méthode de classification qui est la méthode de K-plus

proches voisins (Kppv) pour associer une ou plusieurs catégories à un document non classé.

### III. Description des approches implémentées

#### • Première approche

La première approche utilise les deux méthodes de la représentation des textes. Elle commence par la représentation « sac de mots » dont les documents seront représentés par des vecteurs de mots ensuite, elle utilise la représentation conceptuelle qui est basée sur WordNet pour traiter les documents classés et non classés. Par la suite, elle utilise la méthode Kppv pour classer les documents non classés. La figure 3.1 illustre le processus implémenté.

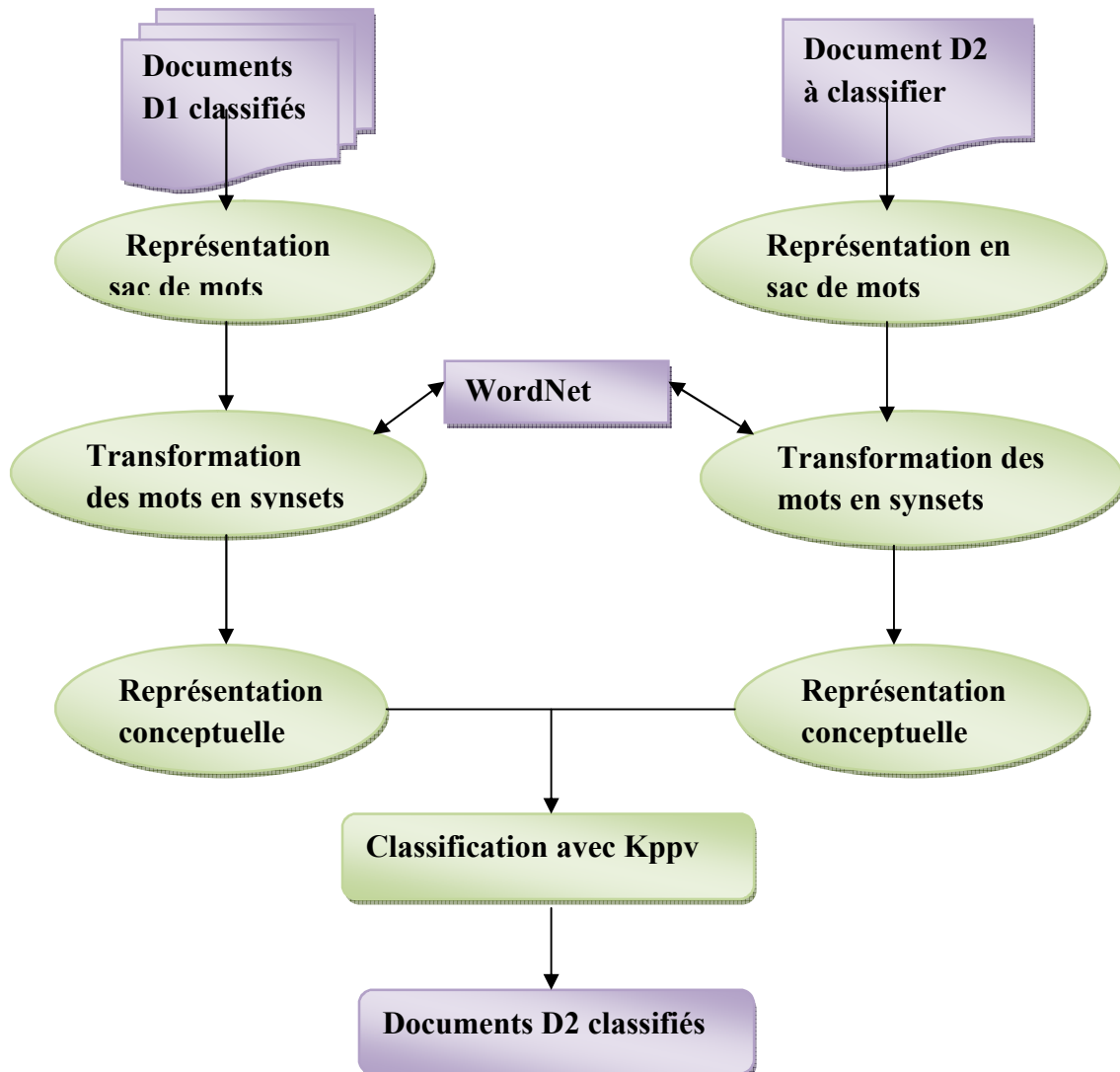


Figure 3.1: Processus de la première approche

**• Deuxième approche**

Dans cette approche nous avons décidé d'introduire les mesures de similarités sémantiques afin d'essayer d'améliorer les résultats de la catégorisation. Des mesures qui servent à prendre en compte les relations sémantiques tels que la synonymie, l'antonymie, père de, fils de...

Tous ce plus que nous avons ajouté par rapport à la première approche nous le mettons dans un algorithme qui est l'algorithme d'enrichissement que nous allons le voir par la suite.

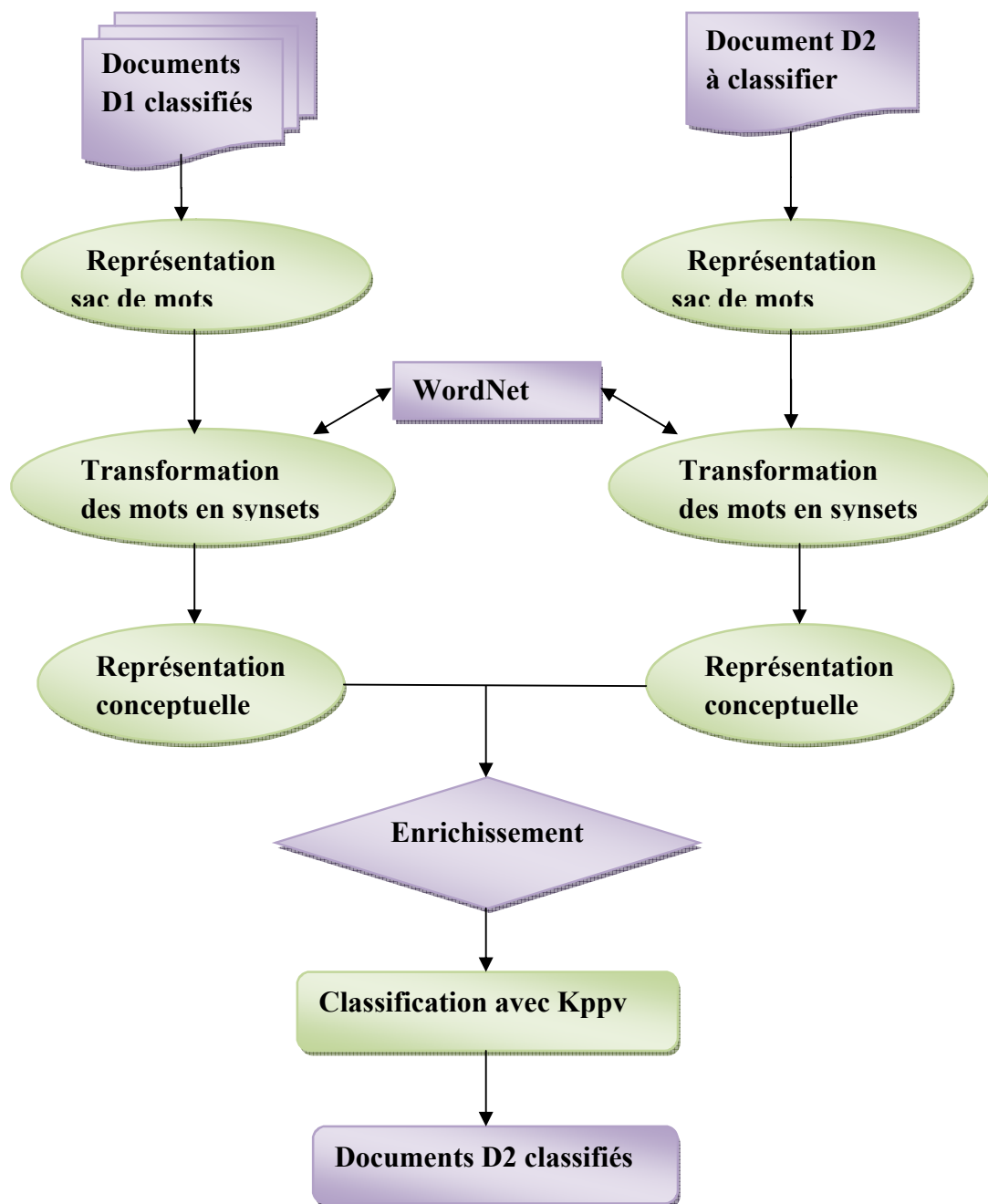


Figure 3.2 : Processus de la deuxième approche

### III.1 Représentation en sac de mots

La première étape utilise la représentation en sac de mots qui consiste à mettre en œuvre une série de prétraitements sur les documents classés et les documents non classés pour extraire l'ensemble des mots, les textes sont transformés en vecteur dont chaque composante représente un mot. Le prétraitement consiste à :

- ✓ Convertir les majuscules du texte en minuscules.
- ✓ Enlever toute la ponctuation du texte
- ✓ Eliminer les mots vides (pronoms personnels, prépositions,...).
- ✓ Mettre les mots du texte dans un vecteur.

### III.2 Transformation des mots en synsets

Après la représentation de chaque document par son vecteur, nous passons à l'étape de la transformation des mots en synsets et cela grâce à une base lexicographique WordNet dans laquelle les mots sont regroupés au sein de groupes de synonymes appelés synsets qui indique un sens différent du mot. La base lexicographique WordNet renvoie une liste ordonnée de synsets pour chaque mot qui peut ajouter le bruit à la représentation et peut induire une perte d'information. Alors il y aura un problème de la désambiguïsation de sens. Pour éviter ce problème, notre processus consiste à remplacer directement chaque mot par sa première signification en considérant qu'elle est la plus appropriée. La figure ci-dessous montre qu'un mot peut avoir plusieurs sens.

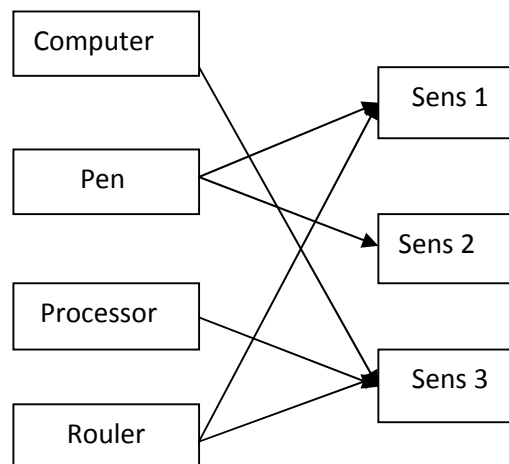


Figure 3.3 : Combinatoire des sens

- Exemple de groupes de synset :

- **Noun :**
  - ✓ **BICYCLE:** [Synset: [Offset: 2734941] [POS: noun] Words: bicycle, bike, wheel, cycle -- (a wheeled vehicle that has two wheels and is moved by foot pedals)].
  - ✓ **HUMAN:**[Synset: [Offset: 6026] [POS: noun] Words: person, individual, someone, somebody, mortal, human, soul -- (a human being; "there was too much for one person to do")]
- **Verb:**
  - ✓ **MAKE :** [Synset: [Offset: 2484888] [POS: verb] Words: make, do -- (engage in; "make love, not war"; "make an effort"; "do research"; "do nothing"; "make revolution")]
  - ✓ **WRITE:** [Synset: [Offset: 1649807] [POS: verb] Words: write, compose, pen, indite -- (produce a literary work; "She composed a poem"; "He wrote four novels")]
- **Adverb:**
  - ✓ **ALWAYS:** [Synset: [Offset: 19245] [POS: adverb] Words: always, ever, e'er -- (at all times; all the time and on every occasion; "I will always be there to help you"; "always arrives on time"; "there is always some pollution in the air"; "ever hoping to strike it rich"; "ever busy")]
- **Adjectif:**
  - ✓ **SMALL:**[Synset: [Offset: 1343705] [POS: adjective]Words: small, little -- (limited or below average in number or quantity or magnitude or extent; "a little dining room"; "a little house"; "a small car"; "a little (or small) group"; "a small voice")]

### III.3 Représentation conceptuelle

La représentation conceptuelle se base sur le formalisme vectoriel pour représenter les documents. Les éléments de cette représentation ne sont plus des mots, mais plutôt des concepts, et cela grâce à l'étape précédente qui est la transformation des mots en synsets. Le schéma ci-dessous illustre la représentation matricielle d'un corpus où les lignes représentent les documents du corpus, les colonnes représentent les termes(les concepts), et l'intersection entre un document  $D_i$  et un terme  $T_j$  représente le nombre d'occurrences du terme  $T_j$  dans le document  $D_i$ .

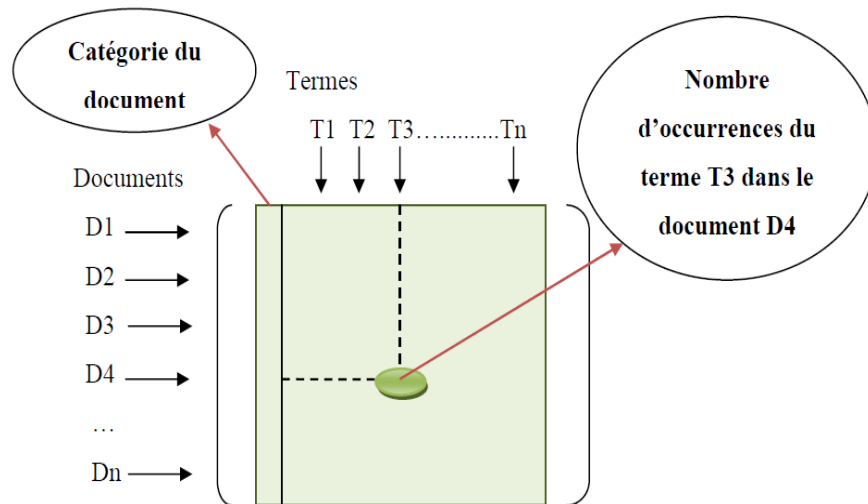


Figure 3.4 : Représentation matricielle d'un corpus [38]

### III.4. Enrichissement

L'enrichissement est une étape très importante dans notre travail. Introduite dans la deuxième approche, elle consiste à calculer la mesure de similarité sémantique entre les concepts qui se retrouvent dans les documents à classer, et chaque concept appartenant au « dictionnaire des concepts ».

Dans notre processus, un dictionnaire est obtenu lors de la phase d'apprentissage et contient les concepts extraits des documents classifiés. Le calcul de mesure de similarité sémantique intervient au cas où le concept extrait du document à classer n'existe pas dans le dictionnaire des concepts dont le but est de le remplacer par le concept qui lui est le plus proche. En effet, après avoir calculé toutes les mesures de similarité sémantique du concept  $C_t$  à classer avec tous les concepts du dictionnaire, le concept  $C_t$  sera remplacé par le concept dont la valeur de la mesure de similarité est la plus grande et dont la valeur dépasse un seuil donné. Dans notre cas, nous avons choisi un seuil de 0.7.

Ainsi le poids du concept élu sera incrémenté dans le tableau de la représentation matricielle. L'algorithme suivant résume les étapes du processus d'enrichissement.

---

**Les entrées :** le dictionnaire de concepts  $D_c$ , le concept  $C_t$  absent du dictionnaire.

**Les paramètres :** seuil et mesure de similarité sémantique.

**Début :**

Indice=0 (l'indice du concept choisi)

Maximum=0 (la valeur maximale de la mesure de similarité)

Pour chaque concept  $C_i \in \text{Defaire}$

Calculer la mesure de similarité sémantique avec Ct ( $\text{Sim}(\text{Ct}, C_i)$ ).

Si ( $\text{Sim}(\text{Ct}, C_i) > \text{maximum}$ ) et ( $\text{Sim}(\text{Ct}, C_i) > \text{seuil}$ ) Alors

Indice= i

Maximum= $\text{Sim}(\text{Ct}, C_i)$

Fin si

Fin pour

remplacer le concept Ct par le concept  $C_{\text{indice}}$  et mettre à jour son poids dans le vecteur du document à classifier

**Fin**

• **Exemple de l'étape d'enrichissement :**

Etant donné le dictionnaire des concepts(2734941, 6026, 19245,2484888, 1649807, 1343705) et le mot « car » qui se trouve dans le document D1 à classifier, nous remarquons que le concept contenant le mot« car » n'existe pas dans le dictionnaire alors nous exécutons l'algorithme d'enrichissement.

Le tableau ci-dessous présente les valeurs des mesures de similarité sémantique calculées

Type	Noun	Noun	Adverb	Verb	Verb	Adjectif
offset	2734941	6026	19245	2484888	1649807	1343705
Mesure de Wu Palmer	0.81	0.5	–	–	–	–

Nous avons la plus grande valeur est celle du premier concept et elle dépasse le seuil. Dans ce cas la le tableau des poids des concepts à classifier devient :

	2734941	6026	19245	2484888	1649807	1343705
<b><u>D1</u></b>	Poids+1	poids	poids	poids	poids	Poids



### III.5. Classification avec Kppv

Après l'enrichissement, nous allons passer à la classification des documents avec la méthode des K plus proches voisins qui compare les textes les plus proches sémantiquement au document à classer pour les regrouper par catégorie. Nous avons choisi cet algorithme pour sa simplicité et sa fréquente utilisation dans le domaine de la catégorisation des textes. L'algorithme de catégorisation de K-plus proches voisins pris de [Jalam, 2003], est le suivant :

---

**Paramètre** : le nombre K de voisins

**Contexte** : un échantillon de L textes classés en  $C = c_1, c_2, \dots, c_n$  classes

**Début**

Pour chaque texte T faire

Transformer le texte T en vecteur  $T = (x_1, x_2, \dots, x_m)$ ,

Déterminer les K plus proches textes du texte T selon une métrique de distance,

Combiner les classes de ses K exemples en une classe C.

Fin pour

**Fin**

**Sortie** : le texte T associé à la classe C.

---

Lors de la classification d'un nouveau document à classer dans une ou plusieurs catégories, il sera comparé aux documents classés à l'aide d'une mesure de similarité qui est la mesure de cosinus. Cette dernière est l'une des premières mesures à avoir été utilisée dans le domaine de la recherche d'information.

Rappelons que la mesure de similarité cosinus entre deux documents a et b, est calculée comme suit :

$$\text{Cosinus (a, b)} = \frac{\sum (P_t(a) \times P_t(b))}{\sqrt{\sum P_t(a)^2 \times \sum P_t(b)^2}}$$

Où :  $P_t(a)$  est le poids du terme t dans le document a et  $P_t(b)$  est le poids du terme t dans le document b.

La valeur de K est un paramètre à déterminer lors de l'utilisation de ce type de classificateur.

## IV. Technologies et outils de développement

Dans cette section, nous allons présenter l'environnement de développement qui va supporter notre application ainsi les outils et langages utilisés.

### IV.1. Langage JAVA

Pour le choix de programmation de notre système nous avons opté pour le langage JAVA et cela pour de nombreuses raisons :

- JAVA est un langage orienté objet simple, qui réduit le risque des erreurs d'incohérence.
- Il permet d'accéder d'une manière simple aux fichiers et aux réseaux (notamment Internet).
- Il est caractérisé aussi par la réutilisation de son code ainsi que la simplicité de sa mise en œuvre.
- Il possède une riche bibliothèque de classes comprenant des fonctions diverses telles que les fonctions standards, le système de gestion de fichiers ainsi que beaucoup de fonctionnalités qui peuvent être utilisées pour développer des applications diverses.

### IV.2. Environnement de développement

L'environnement de développement utilisé, est NetBeans 6.1, il possède de nombreux avantages qui sont à l'origine de son énorme succès dont les principaux sont :

- Un environnement de développement intégré (EDI).
- Permet de supporter différents autres langages, comme Python, C, C++, JavaScript, XML, Ruby, PHP et HTML. Il comprend toutes les caractéristiques d'un IDE moderne (éditeur en couleur, projets multi-langage, refactoring, éditeur graphique d'interfaces et de pages Web).
- La construction incrémentale des projets JAVA grâce à son propre compilateur, qui permet en plus de compiler le code même avec des erreurs, de générer des messages d'erreurs personnalisés, de sélectionner la cible, ...

### IV.3. WordNet

Afin d'implémenter notre travail, nous avons utilisé WordNet de version 2.0 qui est une base de données lexicographique. Le choix de WordNet est cause de diverses raisons :

- C'est la base la plus riche et la plus générale qui contient tous les domaines,
  - Il utilise la langue anglaise qui est la langue la plus utilisée dans le monde.
- Des versions de ce dernier existent pour d'autres langues.

La structure du Wordnet repose sur des ensembles de synonymes appelés synset. Chaque synset représente un sens, un concept de la langue anglaise. Chacun d'eux contient tous les mots synonymes pouvant exprimer le sens auquel il fait référence. Les liens sémantiques ne relient alors pas les mots entre eux mais les synsets auxquels les mots sont affectés.

Le tableau ci-dessous montre la structure de WordNet d'anglais noté EWN en nombre de mots, nombre de synsets et nombre de sens, globalement et par catégorie grammaticale. La plupart sont des noms (74.6%), le reste étant constitué par des adjectifs (14.6%), des verbes (7.6%) et des adverbes (3.2%). La polysémie (nombre de sens par mot) se manifeste dans Wordnet par le fait qu'il y a des mots qui peuvent appartenir à plusieurs synsets (146350 formes traitées / 111223 synsets).

<i>Catégorie</i>	<i>Mots</i>	<i>Concepts</i>	<i>Total Paires Mots-Sens</i>
<i>Nom</i>	109195	75804	134716
<i>Verbe</i>	11088	13214	24169
<i>Adjectif</i>	21460	18576	31184
<i>Adverbe</i>	4607	3629	5748
<i>Total</i>	146350	111223	195817

**Tableau 3.1 : Caractéristiques du nombre de mots et de concepts dans WordNet**

#### IV.4. JWNL

JWNL (Java WordNet Library) est disponible sur: (<http://sourceforge.net/projects/jwordnet/>) est une API Java pour avoir accès au dictionnaire relationnel WordNet dans des formats multiples, aussi bien que la découverte des relations hiérarchiques et de traitement morphologique. Elle est compatible avec des versions WordNet 2.0 à 3.0 et est une mise en œuvre Java complète.

#### IV.5. Corpus utilisé

Dans notre travail, nous avons utilisé le corpus Reuters10 qui est un corpus multi-label contenant 10 catégories possibles basé sur les 10 plus grandes catégories de Reuters-21578. Ce dernier est un ensemble de dépêches financières émises au cours de l'année 1987 par l'agence Reuters, en langue anglaise, et disponible gratuitement sur le web. Ce corpus est une mise à jour du corpus Reuters-22173 qui a permis de supprimer les documents présents deux fois, de corriger des erreurs typographiques, de préciser certains formats, et de mieux définir le découpage à considérer pour l'apprentissage et le test.

Classes	Train	Test
acq	1650	719
corn	181	56
crude	389	189
earn	2877	1087
grain	433	149
interest	347	131
money-fx	538	179
ship	197	89
trade	369	117
wheat	212	71

**Tableau 3.2 : Caractéristiques du corpus utilisé**

A partir du tableau 3.2 nous avons pris pour chaque classe de train 100 fichiers et pour celles de test 10 fichiers.

#### V. Evaluation de notre travail

Afin de pouvoir montrer le rôle de l'utilisation des mesures de similarité sémantique pour l'enrichissement dans la classification des documents, nous avons appliqué les approches sur un corpus qui est Reuters 10. Sans oublier, que notre choix pour le classificateur est la méthode des K plus proches voisins, et cela pour la mise en œuvre de notre travail. Pour évaluer nos approches, nous avons choisis des paramètres d'évaluation qui sont la précision (P) et le rappel (R).

- Pour la première approche nous n'avons pas utilisé les mesures de similarité sémantiques, c'est une étape qui nous permet de classer des documents non classifiés selon des concepts qui sont identiques aux concepts des documents classifiés. Les résultats obtenus sont comme suit :

Classes	K=3	
	P	R
Acq	0.9	0.9
Corn	0.4	0.6
Crude	0.63	0.7
Earn	0.69	0.9
Grain	0.3	0.3
Interest	0.5625	0.9
Money-fx	0.6	0.3
Ship	0.8	0.4
Trade	0.5	0.6
Wheast	0.33	0.1

**Tableau 3.3 : Précision et rappel pour la première approche**

Nous calculons le F pour voir le pourcentage de la catégorisation selon les résultats de précisions et rappel.

$$F = \frac{2 * R * P}{R + P} = 0.5706$$

- Pour la deuxième approche, qui se base sur l'étape d'enrichissement nous avons utilisé les mesures de similarité sémantiques afin d'avoir un meilleur classement pour les documents non classés.

Cet enrichissement consiste à calculer des mesures de similarités sémantiques entre concepts afin de pouvoir comparer les concepts qui se rapprochent, on se basant sur un seuil choisi qui est égale à 0.7, ainsi que l'utilisation de différentes mesures de similarité sémantiques :

- Cas1 : Utilisation de la mesure de similarité de Wu Palmer qui est basé sur les arcs.

Classes	K=3	
	P	R

Acq	0.9	0.9
Corn	0.4375	0.7000000000000001
Crude	0.7000000000000001	0.7000000000000001
Earn	0.6923076923076923	0.9
Grain	0.2727272727272727	0.30000000000000004
Interest	0.5625	0.9
Money-fx	0.6000000000000001	0.30000000000000004
Ship	0.8	0.4
Trade	0.5	0.6000000000000001
Wheat	0.5	0.1

**Tableau 3.4 : Précision et rappel pour la mesure de Wu Palmer**

**Fwu=0.58**

- Cas 2: Utilisation de la mesure de similarité de Lin qui est basé sur le contenu informationnel.

Classes	K=3	
	P	R
Acq	0.9	0.9
Corn	0.4	0.6000000000000001
Crude	0.6363636363636364	0.7000000000000001
Earn	0.6923076923076923	0.9
Grain	0.30000000000000004	0.30000000000000004
Interest	0.5625	0.9
Money-fx	0.75	0.30000000000000004
Ship	0.6666666666666666	0.4
Trade	0.5	0.6000000000000001
Wheast	0.3333333333333333	0.1

**Tableau 3.5 : Précision et rappel pour la mesure de Lin**

**Flin=0.5712**

**VI. Discussion**

Dans la section précédente, nous avons évalué nos approches implémentées pour les comparer afin d'avoir une meilleure catégorisation des textes. Les tableaux ci-dessus, montrent respectivement les résultats de la première approche et les résultats de la deuxième approche, ou nous avons introduit l'étape d'enrichissement qui comprend le calcul des mesures de similarité de Wu Palmer et de Lin sur le corpus Reuters10. Sur cette collection et dans la majorité des cas, les résultats de la deuxième approche sont légèrement supérieurs à la première approche. Donc l'étape de l'enrichissement nous a permis d'avoir une amélioration dans la catégorisation des textes.

**VII. Conclusion**

Ce chapitre présente la description et la mise en œuvre des étapes implémentées pour notre approche, qui avaient pour intérêt d'utiliser les mesures de similarité pour la catégorisation automatique des textes et cela, en utilisant une représentation conceptuelle basée sur WordNet ainsi qu'un corpus Reuters 10. Enfin nous avons utilisé la méthode Kppv pour attribuer à chaque document sa catégorie.

## **Conclusion générale**

Dans ce mémoire, nous avons développé un travail dans le cadre de l'enrichissement de la représentation conceptuelle dans la catégorisation des textes en utilisant des mesures de similarité sémantique. L'enrichissement élaboré nous a permis d'évaluer la proximité sémantique entre les concepts afin de voir l'amélioration de la performance de notre classificateur.

Les résultats obtenus avec enrichissement nous ont permis de remarquer une légère amélioration par rapport aux résultats obtenus sans enrichissement. L'enrichissement a été effectué en utilisant deux mesures de similarité sémantique à savoir la mesure Wu Palmer et la mesure Lin. La représentation conceptuelle a été obtenue en utilisant la base lexicographique WordNet 2.0. L'évaluation des approches se basaient sur un échantillon du corpus Reuters-Top 10.

Malheureusement, le temps est court et il a été nécessaire d'ajouter d'autres mesures de similarité sémantique, de fixer certains paramètres pour en étudier d'autres plus en profondeur ainsi que plusieurs seuils. Évidemment, il aurait été intéressant d'observer le comportement de nos approches implémentées sur d'autres corpus plus riches, ainsi que sur d'autres classificateurs. Notre perspective dans un premier temps est de consolider la démarche implémentée en évaluant sur d'autres collections, puis élargir notre domaine en ajoutant d'autres mesures de similarité sémantique et aussi de travailler avec la dernière version de WordNet 3.0.



## Références bibliographiques

- [1] Jalam, Radwan "*Apprentissage automatique et catégorisation de textes multilingues*". Thèse de doctorat, Université Lumière Lyon 2, France, juin 2003.
- [2] FABRIZIO SEBASTIANI « Machine learning in automated text categorization », Conseil recherché National, Italie, Mars 2002.
- [3] Dictionnaire de l'Académie Française CINQUIÈME ÉDITION, 1798.
- [4] Dictionnaire de l'Académie française, 8e édition ,1992.
- [5] M. F. Porter. "*An algorithm for suffix stripping*". Program, 1980.
- [6] Camelia Ignat, « Représentation de textes à l'aide d'étiquettes sémantiques dans le cadre de la classification automatique », European Commission, IPSC, Strasbourg France, 2007.
- [7] Jalam, R. and Chauchat, J.-H. « *Pourquoi les n-grammes permettent de classer des textes? Recherche de mots-clés pertinents à l'aide des n-grammes caractéristiques* ». In Morin, A. and Sébillot, P., editors, 6èmes Journées internationales d'Analyse statistique des Données Textuelles, St. Malo France. 2002.
- [8] Dunning, T. « *Statistical Identification of Languages* ». Technical Report MCCS 94-273, Computing Research Laboratory .1994.
- [9] Simon RÉHEL, « Catégorisation automatique de textes et Cooccurrence de mots provenant de documents non étiquetés », Mémoire, Université Laval Québec, Canada, Janvier 2005.
- [10] V. Vapnik « The Nature of Statistical Learning ». 1995.
- [11] Romain VINOT, « *Classification automatique de textes dans la catégorie non thématique* », thèse pour obtenir le grade de docteur d'école nationale supérieure des télécommunications, Présentée et soutenue le 09 février 2007. 2004.
- [12] J. Rocchio « Relevance feedback in information retrieval », 1971.
- [13] Karima ABIDI, « La catégorisation de texte Multilingue », Mémoire de magistère Ecole supérieure d'Informatique, Algérie, 2010-2011.
- [14] Grzegorz DZICZKOWSKI, « *Analyse des sentiments : système autonome d'exploration des opinions exprimées dans les critiques cinématographiques* » thèse de doctorat soutenus le 4 Décembre 2008.
- [15] R. Baeza-Yates et B. Ribeiro-Neto. *Modern Information Retrieval*. ACM Press; Addison-Wesley: New York; Harlow, England; Reading, Mass., 1999.

- [16] G. Salton et M. J. McGill, Introduction to modern information retrieval. McGraw-Hill. New York, 1983.
- [18] Z. Wu et M. Palmer. Verb semantics and lexical selection. In Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics, 1994.
- [19] D. Lin. An Information-Theoretic Definition of similarity. In Proceedings of The Fifteenth International Conference on Machine Learning (ICML'98) Morgan-Kaufmann: Madison, WI, 1998.
- [20] R. Rada, H. Mili, E. Bichnell et M. Blettner, Development and application of a metric on semantic nets. IEEE Transaction on Systems, 1989.
- [21] 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.
- [22] Lexical chains as representations of context for the detection and correction of malapropisms. In *Christiane Fellbaum (editor), WordNet: An electronic lexical database, Cambridge, MA: The MIT Press.* 1998.
- [23] Zargayouna, H., Salotti, S.: Mesure de similarité dans une ontologie pour l'indexation sémantique de documents XML. Actes de la conférence IC'2004. 2004.
- [24] P. Resnik. Using information content to evaluate semantic similarity in taxonomy. In Proceedings of 14<sup>th</sup> International Joint Conference on Artificial Intelligence, Montreal, 1995.
- [25] D. Lin. An Information-Theoretic Definition of similarity. In Proceedings of The Fifteenth International Conference on Machine Learning (ICML'98). Morgan-Kaufmann: Madison, WI, 1998.
- [26] G. A. Miller, R. Beckwith, C. Fellbaum, D. Gross, et K. Miller. Introduction to WordNet: An On-line Lexical Database. Cognitive Science Laboratory, Princeton University, Princeton, Technical Report 1993.
- [27] Seco, N., T. Veale, et J. Hayes. An intrinsic information content metric for semantic similarity in wordnet. *Proceedings of ECAI'2004, the 16<sup>th</sup> European Conference on Artificial Intelligence, Valence, Espagne.* 2004.
- [28] 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.

- [29] C. Leacock et M. Chodorow. Combining Local Context and WordNetSimilarity for Word Sense Identification. In *WordNet: An Electronic LexicalDatabase*, C. Fellbaum, MIT Press, 1998.
- [31] PIRRÓ, G. et EUZENAT, J. A feature and information theoretic framework for semantic similarity and relatedness. *In ISWC 2010*.
- [32] Amos Tversky. Features of similarity. In *Psychological Review*, 1977.
- [33] P. Siddharth, S. Banerjee et T. Pedersen. Using measures of semanticrelatedness for word sense disambiguation. In *Proceedings of the Fourth International Conference on Intelligent Text Processing and Computational Linguistics*, Mexico City, Mexico. 2003.
- [34] I.Gurevych et M. Strube. Semantic similarity applied to spoken dialoguesummarization. In *Proceedings of the 20th International Conference onComputational Linguistics*, Geneva, Switzerland, 2004.
- [35] G. Hirst et A. Budanitsky Correcting real-word spelling errors by restoring lexical cohesion. *Natural Language Engineering* 2004.
- [36] W. Lord, R.D. Stevens, A. Brass et C.A.Goble. Semantic SimilarityMeasures as Tools for Exploring the Gene Ontology. *Pacific Symposium onBiocomputing* 8, 2003.
- [37] H. Jeffrey, L. William et D. John. A Semantic Similarity Measure for SemanticWeb Services. In *proceedings of WSS05*. 2005.
- [38] TERKIA DERDRA Amel, BENSFIA Fatima Zahra, *La Représentation Conceptuelle pour laCatégorisation des Textes Multilingue*, 2012.

## Résumé

Ce mémoire s'inscrit dans la problématique générale liée à l'évaluation de l'utilisateur des mesures de similarité dans la classification des textes. Le but est de représenter tous les documents sous forme d'une représentation vectorielle dont les composants seront des termes. Notre objectif est d'associer à chaque document non classé sa catégorie en se basant sur un ensemble de textes préalablement classé et sur des mesures de similarité.

La représentation conceptuelle s'appuie sur des concepts issus de la base de données lexicographique WordNet et l'expérimentation est effectuée sur un corpus extrait du corpus Reuters 10.

**Mots clés :** représentation en sac de mots, représentation conceptuelle, classification des textes, mesures de similarité, WordNet.

## Abstract

This memory falls under the general problems related to the evaluation of the user of similarity measurements in the texts classification. The goal is to represent all the documents in the form of a vectorial representation whose components will be terms. Our objective is to associate with each document not classified its category while being based on a whole of texts classified beforehand and to measures of similarity.

The conceptual representation is based on concepts resulting from lexicographical data base WordNet and the experimentation is carried out on a corpus extracted the corpus Reuters 10.

**Key words:** representation out of bag of words, conceptual representation, texts categorization, similarity measurements, WordNet

## ملخص

تصنيف التشابه تدابير بتقييم المتعلق العام للمشكلة نتحدثنا المذكرة هذه النص. الشكلية نفس باستخدام الفئات والوثائق جميع تمثيل هو الغرض. ذلك من هدفنا عن سابقا المصنفة النصوص من مجموعة إلى استنادا مصنفة غير وثيقة كل بطهو التشابه تدابير طريق. البيانات قاعدة من النظرية المفاهيم على يستند التمثيل هذا المعجمية wordNet منظمة وكالة الاخبار العالمية من مستخرجة وثائق قاعدة على التجربة تنفيذ يتمو

Reuter top 10

**المفتاح:** الكلمات, تدابير تصنيف النصوص, المفاهيم تمثيل

التشابه, WordNet