Table des matières

Liste des figures

Liste des tableaux

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 àl'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 $(_{dj,ci})$ □ D × C, où D l'ensemble des textes et C l'ensemble des catégories. La valeur V (Vrai) est alors associée au couple $(_{dj,ci})$ si le texte dj appartient à la classe ci 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 : Φ: D × C → {V, F} qui associe une ou plusieurs étiquettes (catégories) à un document $_{dj}$ telle que la décision donnée par cette procédure coïncide le plus possible avec la fonction $^{\Phi^{\wedge}}$: D × C → {V, F}, la vraie fonction qui retourne pour chaque vecteur dj 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 classifier 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 définit 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 représentions des textes est la phase la plus importante dans le processus decatégorisation des textes, et cela pour classifier 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 documentLes 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,dontlescomposantessont les mots contenus dans le texte. L’analyselexicale estunprocessus qui permet deconvertir le texte d’un documenten un ensemble demotset qui permet de reconnaitre les espaces deséparations des mots, le signes deponctuations,…etc. ; ces dernierssontsupprimésde la représentation.

L'avantage de cette représentation estd'exclure toute analysegrammaticale ettoute notion dedistanceentre lesmots, mais l'inconvénient est que le regroupement des motsdudocumentsans prendre en compte les combinaisonset l'ordre des mots dans laphrase entraine une perte dans la sémantique detexte.

b. Représentation avec les racines lexicales

Cette représentation consiste àregrouper les mots de la mêmeracine dans une seulecomposante, alors que cetteracine peut être commune pour des motsqui ontdes sens différents. On a parexemple les mots «descendre»«descendance » et«descendu» ont la même racine«descend » mais leurs notions sans différentes. Cette méthode se faitgrâce à des algorithmes qui ont été proposés; l'un desplus 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'ellenécessite uneanalysegrammaticale des textes afin de remplacer tous lesverbes du document parleur formeinfinitive et les noms par leur formeau singulier. L'objectif de cette représentation estd'associer à chaquemot, uneentrée dans le lexique qui est définitcomme un ensemble delemmes. Parexemple le lemme de « chantaient » est« chanter ». Cetteméthode est simplemais elle peut causer plusieurs problèmescomme La perte de l'informationdonnée par lecontexte syntagmatique, nécessaireà la distinction des lemmespolysémiques et La présencede synonymes,considéréscomme des lemmesdifférents même s'ils font référence aumêmeconcept[6].

d. Représentation avec les n-gramme

Cette représentation consiste à représenter le document par des n-grammesqui sontuneséquence de n caractères consécutifs. Plusieurstravaux ont montré l'efficacitédes n-grammescomme méthode dereprésentation des textes pour la classification. Cette méthode à plusieursavantages, comparativement à d'autres techniques, Les n-grammescapturentautomatiquement les racines des mots les plusfréquents sanspasser par l'étape de recherchedes racines lexicales[7].Ilintroduit aussi la notion d'indépendancede lalangue comme montré dans[8]. Les espaces sont pris enconsidération parce qu'en effet, la non prise encomptede ces derniersintroduitdu bruit.

e. Représentation conceptuelle

Cette étape sert à représenter le document textuel sous forme d'unensembledeconcepts qui est un groupe de synonymes appelés *«synsets»*.Selon Rehel dans [9],l'avantage de la représentation conceptuelle est deréduirel'espace de travail car les mots qui sont synonymes partagent aumoins unconcept.L'inconvénient decette représentation est qu'il n'existepas 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 undocument. L'objectif est de trouver les termes qui représentent le mieux lecontenud'un document. Les méthodes les plus populaires sont :

a. Mesure TF (Term Frequency)

Ce facteur permet de mesurer la représentativité locale d'un terme. Ilprend en compte les informations locales du terme qui ne dépendent quedu document donné,et indique l'importance du terme dansce document.

b. Mesure TFIDF (Term Frequency Inverse Document Frequency)

La mesure *Tf-Idf* est une bonne approximation de l'importance d'untermedansundocument, particulièrement dans des corpus de documentsde tailleshomogènes. Le poids d'un terme *T* dans un document *D* est calculécomme suit :

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

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 nécessairepourtraiter des corpus de grandes tailles, donc il est difficile de prendre l'ensemblede tous lesmots comme étant des attributs, car cela engendre une perte de mémoireet de temps decalcul. Pour cela il existe une approche qui est la réduction desdimensions permettant de résoudre ce problème. Son objectif est de

sélectionner oud'extraire un sous-ensemble optimal de caractéristiques pertinentes pour un critèrefixé. Les techniques de réductionde la taille du vocabulaire sont classées en deuxcatégories :

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

Une des méthodes la plus simple et la plus évidente consiste à prendre lesattributs d'origine, conserver seulement ceux utiles à la catégorisation selonune fonction d'évaluation et éliminer des éléments du vocabulaire qui nenous 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, diffèrent types de classificateursont étémis au point.Nous présentons ci-dessous quelques algorithmesd'apprentissagecouramment utilisés dans ce cadre :

a. Machine à support vectoriel (SVM)

Ces machines, proposées par Vapnik dans[10]ont été utilisées avec succèsdansplusieurs tâches d'apprentissage. Le but de SVM est de trouver unclassificateur quisépare au mieux les données etmaximise la distance entreces deux classes. AvecSVM, ce classificateur est un classificateur linéaireappelé hyperplan. Dans le cas de la catégorisation des textes, les entrées sontdes documents et lessorties sont des catégories. SVM estconsidéré parmi les plusperformants pour la catégorisation enraison de samodélisation simpliste et rapide à calculerpar une machine étant donnéqueSVM est un classificateur linéaire. Par contre elle introduit des conceptscomplexes peu adaptés aux corpus de grandes tailles non fixes, sans oublier derappeler 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 aprouvé son efficacité face au domaine de la catégorisation des textes. Sesperformances la

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

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

- **Mesure Cosinus**

Le cosinus est l'une des premières mesures à avoir été utilisée dansledomainede la recherche d'information. Le cosinus entre deuxvecteursestobtenu encalculant le produit scalaire entre ces deuxvecteurs, quenous divisons par leproduit de la norme des deuxvecteurs. Le cosinusentre deuxvecteurs a et b est défini par l'équationsuivante :

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

c. Méthode de Rocchio

Le classificateur de Rocchio parue dans [12], est l'un des plussimples etplus ancien algorithmes de classification du modèle vectoriel. Ceclassificateur a étélargement utilisé dans la catégorisation textuelle.L'avantage de ce type declassificateur est la simplicité et l'interprétabilité. L'apprentissage de ce type declassificateur est souvent précédé par une sélection et uneréduction de termes.

d. Naïve bayes

L'algorithme Naïve Bayes (NB), est une méthode très connue enapprentissage, ellepermet de calculer les probabilités conditionnelles. Dans lecas de la classification destextes,on cherche la classification qui maximise laprobabilité d'observer les mots du document. Lorsde la phase d'entraînement,le classificateur calcule les probabilitésqu'un nouveau documentappartienne àtelle catégorie à partir de la proportion desdocuments d'entraînementappartenant à cette catégorie. Il calcule aussi la probabilité

qu'un mot donnésoit présentdans un texte, sachant que ce texte appartient à tellecatégorie.

Quand un nouveau documentdoit être classé, on calcule lesprobabilitésqu'ilappartienne à chacune des catégories à l'aide de la règle deBayes.**[9].**

e. Les arbres de décision

Les arbres de décision sont composés d'une structure hiérarchique en formed'arbre.Un arbre de décision est un graphe orienté, sans cycles, dont les nœudsportent une question, lesarcs des réponses, et les feuilles des conclusions.Un classificateur de texte basé sur la méthoded'arbrededécision est unarbre de nœuds internesqui sont marqués pardes termes, lesbranches qui sortent desnœuds sont des tests sur les termes, etles feuilles sontmarquées par catégories[13].Une méthode poureffectuerl'apprentissage d'un arbre de décision pour une catégorieCi consiste à vérifiersi tous les exemples d'apprentissage ont la même étiquette. Dansle cascontraire, nous sélectionnons un terme Tk, et nous partitionnonsl'ensembled'apprentissage enclasses de documents qui ont la même valeurpour Tk, et àlafin nouscréons les sous-arbres pour chacune de ces classes.Ce processus est répétérécursivement sur les sous-arbres jusqu'à ce que chaquefeuillede l'arbregénéré de cette façon contienne des exemples d'apprentissage attribuésà lamême catégorie Ci,qui est alors choisie comme l'étiquette de lafeuille.

f. Les réseaux de neurone

Les approches neuronales furent les premières à être utilisées afin de réaliserun apprentissage de type statistique grâce à leur capacité de classification et degénéralisation.Un réseau de neurones artificiels est composé d'une ou deplusieurscouches se succédant dontchaque entrée est la sortie de la couche quila précède.Chaque couche i est composée de Ni neurones, prenant leursentréessur les Ni-1 neurones de la couche précédente. Nous pouvons ainsivoir unréseau de neuronesartificiels comme un réseau ou graphe orienté dontles nœudssont les neuronesartificiels. Le but va être d'attribuer des poidssynaptiques à chaque neuroneafin 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'utiliseruneméthoded'évaluation. Cette dernière doit comporter la précision et le rappel.La précision est défini comme la probabilité conditionnelle ; le rappel mesure lalargeurde l'apprentissage et correspond à la fraction des documents pertinents, parmi ceuxfournis 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'ontpas été.

V. Les applications de la catégorisation des textes

La catégorisation des textes est utilisées 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 fonctiond'une action, d'une stratégie ou d'un processus de raisonnement, etc. Ilspeuvent êtreabstraits ou concrets, élémentaires ou composés, réels ou fictifs.Habituellement, les concepts sont organisés en taxonomie. Une taxonomie

est une hiérarchie deconcepts (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 typed'interactionentre les concepts d'un domaine, elles sont formellement définies comme tout sous-ensemble d'un produit de n ensembles, c'est-à-dire

R: $C_1xC_2x \ldots x\ C_n$. Des exemplesde 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 $n^{ième}$ élément de larelationest unique pour les *n-1* précédents.

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

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 leurinterprétation. Ils sont définis à l'aide de formules bien formées de la logique dupremier ordre enutilisantles prédicats de l'ontologie.

III.5 Les instances

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

206 et *Atos*sontdes 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 ≡ B alors d (A, B) =0 : la distance entre deux concepts équivalents est nulle.
- Si A ⊑ B ⊑ C alors d (A, B) ≤ d (A, C) : la distance entre un concept avec un subsument direct est inferieure à sa distance avec n'importe quel autre subsument.
- Si A ∩ B alors d (A, B) = ∞ : 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.

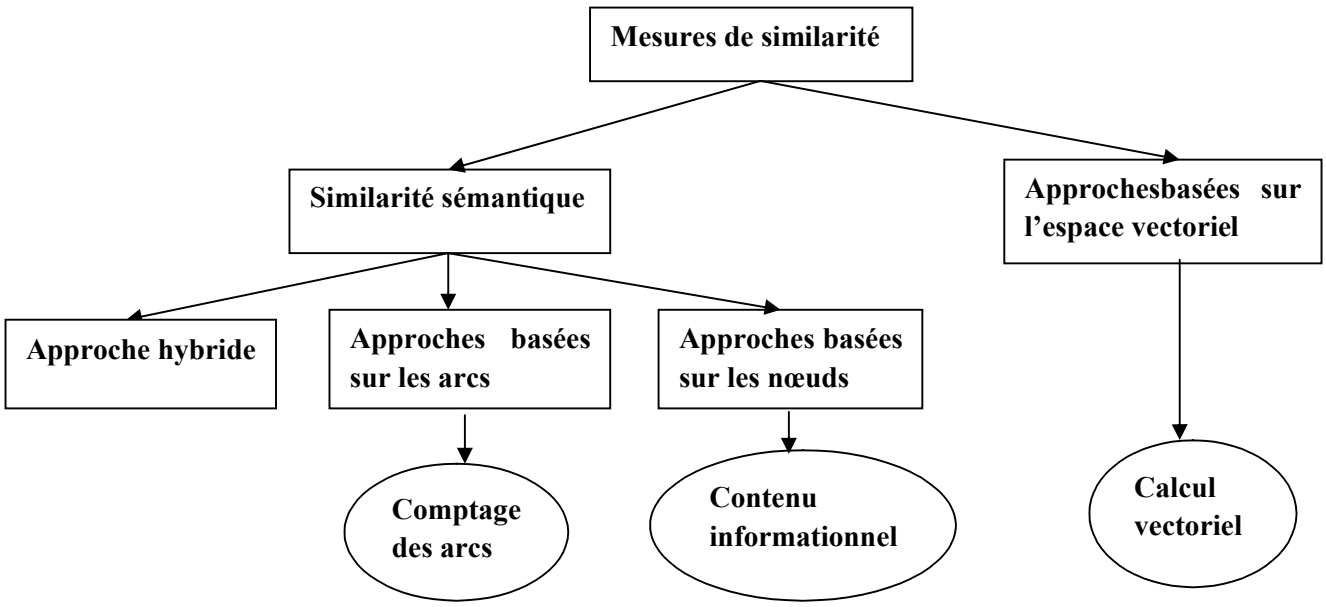


Figure 2.1 :Taxonomie des approches de mesure de similarité.

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

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

a. Mesure de Wu & Palmer

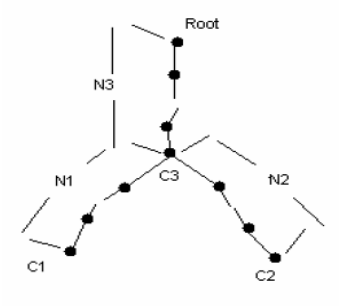
Dans une ontologie, la similarité est définie par rapport à la distance quisépare deux concepts dans la hiérarchie et également par leur position par rapport à laracine.

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

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

Où cest le concept le plus spécifique qui subsume les deux concepts $c_1$et c_2 c'est-à-direle PPG qui sépare c1 et c_2,prof(c)est le nombre d'arcs qui sépare c de la racine etdist (c_i, c)le nombre d'arcs qui séparent $c_i$de c.

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



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

b. Mesure de Rada et al.

Cette mesure de similarité [20] dans un réseau sémantique peut êtrecalculéeense basant sur les liens hiérarchiques «is-a». Pour calculer lasimilarité entre deuxconcepts dans une ontologie, on doit calculer la distanceentre les concepts c'est-à-dire le nombre des arcs minimums qui les séparentpar le chemin le plus court. C'estun moyen pourévaluer la similaritésémantique dans une ontologie hiérarchique.

$$sim_{rada}(c_1, c_2) = \frac{1}{1 + dist(c_1, c_2)}$$

La mesure de [20] utilise une métrique dist (c_1; c_2), qui indique lenombred'arcsminimum à parcourir pour aller d'un concept c1 à un concept c_2.

c. Mesure de Ehrig et al.

Ehrigintroduit dans [21]un travail de mesure de similarité pour les ontologies. Cetravailintroduit trois couches : les données, l'ontologie et le contexte. Lasimilarité des entitésest mesurée au niveau des données en considérant lesvaleurs de donnéesde type simpleou complexe (entiers, caractères). Lesrelations sémantiques entre lesentités sont mesurées au niveau de la couche del'ontologie. Finalement la couche ducontextespécifie comment les entités del'ontologie sont utilisées dans un certaincontexte externe, plus spécifiquement,le contexte de l'application.

d. La mesure de Hirst-St.Onge

La mesure deHirst-St.Onge dans[22] prend en considération toutes lesrelationsdans WordNet. Les liens sont classés comme *haut* (eg. partie-de), *bas*

(eg. sous-classe), *horizontal* (eg. antonyme). La similarité est calculée entremots par le poidsdu chemin leplus court qui mène d'un terme à un autre. Il estcalculé en fonctions deces classificationsqui indiquent les changements dedirection :

$$Sim (c1, c2) = T – chemin – K \times d$$

Tels que *T* et *K* sont des constantes, *chemin* est la longueur du chemin le pluscourten nombre d'arcs et *d* est le nombre de changements de direction. L'idéeest quedeux concepts sont proches sémantiquement si leurs synsets sontconnectés par unchemin quin'est pas très long et qui ne change pas souvent dedirection. S'il n'y apas de chemin, lepoids est égal à zéro.

e. La mesure de Zargayouna

La mesure de [23] est inspirée de celle de [18]. Le lien père-filsestainsi privilégié par rapport aux autres liens de voisinage en adaptant lamesure deWu-Palmer.L'adaptation de la mesure est faite au travers de lafonction de calcul dudegré despécialisation d'un concept (spec) qui mesure sadistance par rapport àl'anti-racine.

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

$$spec(c_1, c_2) = prof_b(c) * dist(c_1, c) * dist(c_2, c)$$

Où *profb(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 conceptsde l'ontologie. La similarité est alors calculée à partir de l'information partagée par les concepts. Le contenu informatif est défini par :

$$IC(c) = -\log p(c).$$

Ou nous avons*c* un concept, et *p(c)* la probabilité de lui trouver un de ses descendantsdans le corpus. Parmi les mesures basées sur le contenu informationnel on peutciter :

a. Mesure de Resnik

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

$$Sim(c1, c2) = CI(ppg(c1, c2))$$

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

b. Mesure de Lin

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

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

Cette mesure se base aussi sur le concept d'IC et elle utilise une approchehybride quicombine deux sources de connaissances différentes (Thesaurus, corpus). Les travauxde Mil[26]ont évalué cette mesure à travers uneexpérience qui utilise des sujetshumains pour évaluer la similarité entre

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

c. Mesure de Seco et al.

Cette méthode calcule le contenu informatif des nœuds à partir de WordNet au lieud'uncorpus. Seco et al. [27] utilisent les hyponymes descendants desconceptspour calculer le contenu informatif de ceux-ci.

$$\frac{log(\frac{hypo(c)+1}{max_{wn}})}{log(\frac{1}{max_{wn}})} = 1 - \frac{log(hypo(c)+1)}{log(max_{wn})}$$

Où : *hypo(c)* est le nombre d'hyponymes du concept c ; et max wn : uneconstantequiindique 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 quiest 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 contenuinformationnel du PPG à ceux des concepts dont on cherche la similarité, elleprendenconsidération aussi le nombre d'arcs. La mesure adoptant cetteméthode est baséesur lacombinaison d'une source de connaissance riche(thesaurus) avec une sourcede connaissance pauvre (corpus). Notons que cetteformule est définie par l'inversede ladistance sémantique.

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

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

$$distance\ (c_1, c_2) = CI\ (c_1) + CI\ (c_2) - (2.CI\ (ppg\ (c_1, c_2)))$$

b. Mesure de Leacock et Chodorow

La mesure de Leacock et Chodorow[29] est une mesure basée sur lalongueur du plus court chemin entre les synsets de wordnet. Le plus courtchemin est celuiqui comprend le plus petit nombre de nœuds intermédiaires.

Cette mesure est définiecomme suit :

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

Où length (c_1, c_2) est le plus court chemin entre deux nœuds et D laprofondeurmaximaledans 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 distancetaxonomique (l = N1+N2), la profondeur du concept commun le plusspécifique dans la taxinomie(h = N3) ainsi que la densité sémantique locale (d = IC (lso (s1, s2))),cette dernière étant exprimée en terme de contenuinformationnel. Leur mesure estexprimée par :

Sim Li (s1, s2) = f (f1 (l), f2 (h), f3 (d))

Où f1, f2 et f3 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 decontenuInformationnel.

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

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 dimensionnelpourreprésenter chaque objet et calculent la similarité en se basant sur la mesurede cosine oula distance euclidienne.La définition de la similarité entre deuxvecteurs d'objets est obtenue par leurs contenus internes. Parmi les approchescitées dans la littérature onpeut 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 leursvecteurs sont confondus. Si deux objets ne sont pas similaires, leurs vecteursforment un angle (X, Y) dont le cosinus représente la valeur de la similarité. Laformule est définie par le rapport du produit scalaire des vecteurs x et y et leproduitde la norme de x et de y.

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

b. Similarité de Jaccard

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

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

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

$\|x\|=\Phi_x$ i=1 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 communsmultipliéspar 2 sur le nombre total d'objets. La mesure de Dice est doncdéfinie par la formulesuivante :

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

VI.5 Mesure de similarité intentionnelle

D'un point de vue ensembliste, deux entités sont similaires si le cardinale 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.Tverskyqui a proposé dans [32]la définition suivante d'une similarité entre deux concepts :

$$sim_{tversky}(A,B)=\alpha.comm(A,B)-\beta.diff(A,B)-\gamma.diff(B,A)$$

Ouα, β, γsont des constantes. Si $\beta = \gamma = 0$ et $\alpha = 1$, la similarité entre A et Bcorrespond 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 alorsune mesurede 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 :

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

VII. Domaine d'application

Afin de connaitre l'importance de la mesure de similarité, il est nécessaire de rappelerles 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 lamétrique dela similarité sémantique pour mesurer la similarité sémantique entretous les sens demot d'une paire donnée de mot et les désambiguïser ainsi dans uncontexte donné. Les auteurs du travail [34] ont appliqué les mesures desimilarité sémantique de WordNet pour évaluer la pertinence des expressions. Letravail deHib[35] étudie l'utilité de la similarité sémantique dans le problème de correction d'orthographes, où des erreurs d'orthographes réelles sont détectées etcorrigées automatiquement.

VII.2 Bioinformatique

Le travail deLor[36] s'intéresse à la similarité sémantique entre les protéines,plutôtque les termes de l'ontologie GO, c'est pour cette raison qu'il a combinéentre troismesures de similarités [24][25] [28].

VII.3 Web Services

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

mesuredesimilarité proposée est basée sur l'intuition que les objets similaires partagentlesinformations 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 desmesures du contenu informationnel pour déterminer la similarité conceptuelle. L'autre familled'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 taches 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) pourassocier 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 serontreprésentés 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.

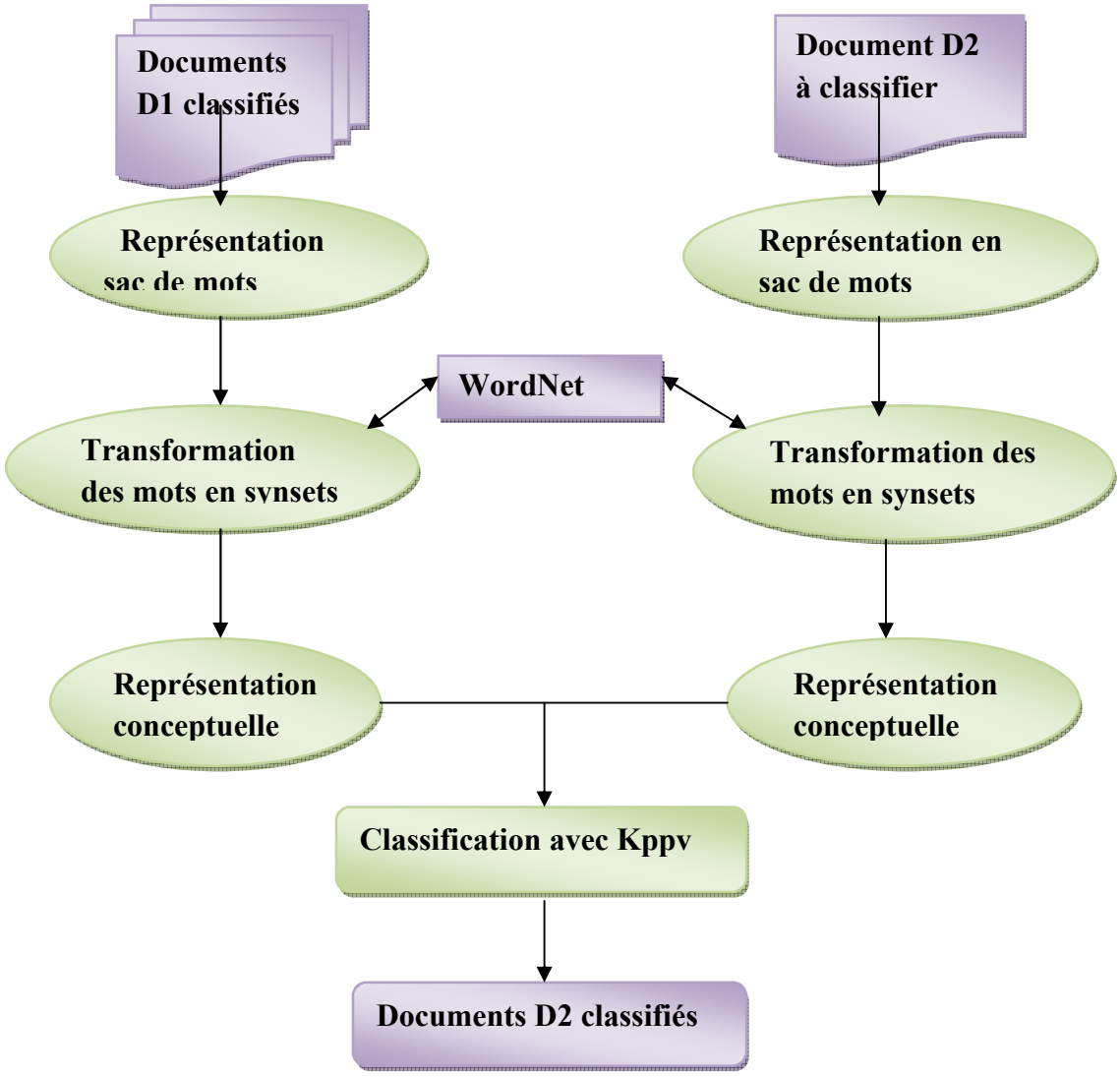


Figure 3.1: Processus de la première approche

- **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.

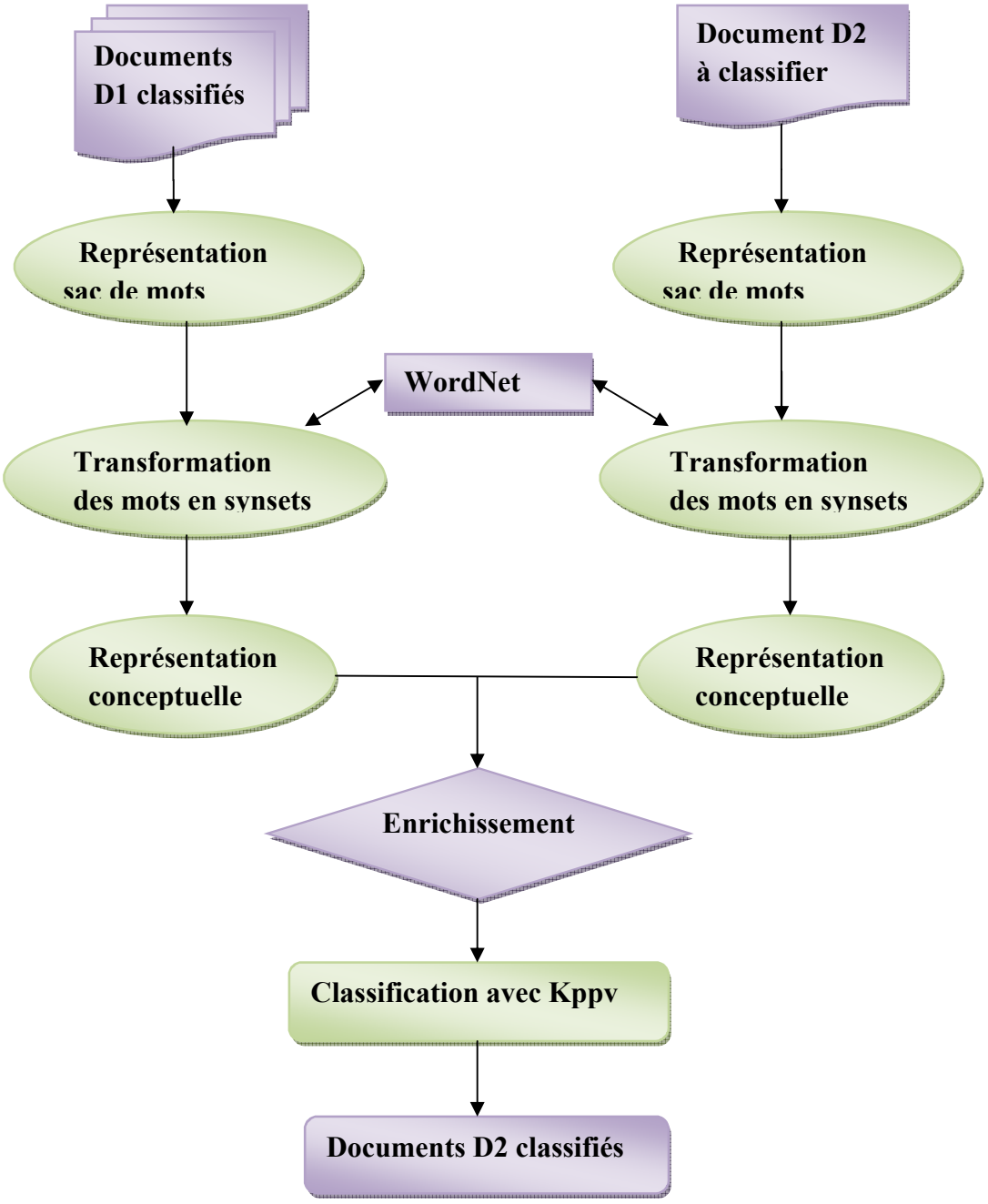


Figure 3.2 : Processus de la deuxième approche

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 désambigüisation 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.

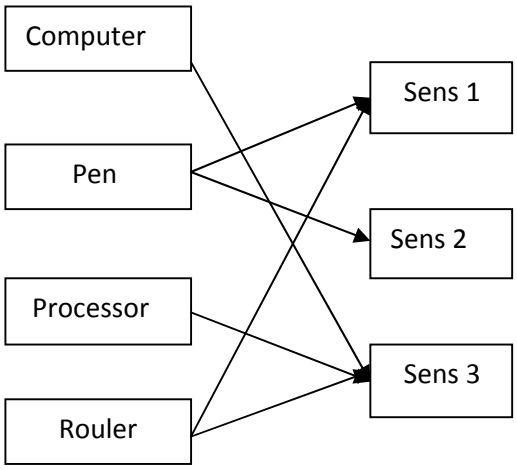


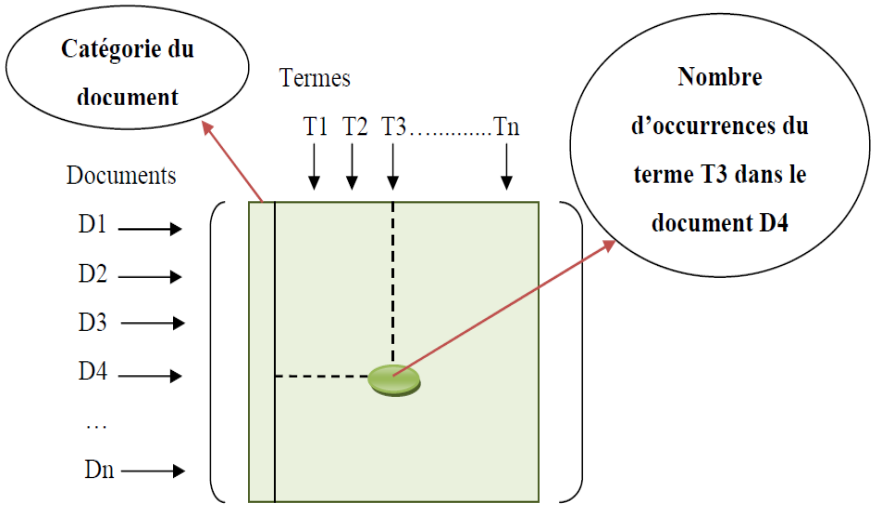
Figure 3.3 : Combinatoire des sens

- **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 Di et un terme Tj représente le nombre d'occurrences du terme Tj dans le document Di.



Figue 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 classifier, 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 classifier n'existe pas dans le dictionnaire des concepts dont le but 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 Ct à classifier avec tous les concepts du dictionnaire.Le concept Ct 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 étapesdu processus d'enrichissement.

Les entrées : le dictionnaire de concepts Dc, le concept Ct 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 $C_i \in Dc$faire

Calculer la mesure de similarité sémantique avec Ct (Sim(Ct,Ci)).

Si (Sim(Ct,Ci) > maximum) et ((Sim(Ct,Ci) > seuil) Alors

Indice= i

Maximum=(Sim(Ct,Ci)

Fin si

Fin pour

remplacer le concept Ct par le concept C_{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
D1	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,…, c_n classes

Début

Pour chaque texte T faire

Transformer le texte T en vecteur T = (x_1, x_2, …, 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 similaritéqui est la mesure de cosinus. Cette dernière est l'une des premières mesures à avoir été utilisée dansledomaine de la recherche d'information.

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

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

Ou : Pt(a) est le poids du terme t dans le document a et Pt(b) est le poids du terme t dansle 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 lelangage 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 fonctionsdiverses 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 denombreux avantages qui sont à l'origine de son énorme succès dont lesprincipaux sont :

- Un environnement de développement intégré (EDI).
- Permet de supporter différents autreslangages, comme Python, C, C++, JavaScript, XML, Ruby, PHP et HTML. Il comprendtoutes 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 messagesd'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 étécause 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. Desversions 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 lalangue anglaise. Chacun d'euxcontient tous les mots synonymes pouvant exprimer lesens auquel il faitréférence.Les liens sémantiques ne relient alors pas les mots entreeux mais lessynsetsauxquels les mots sont affectés.

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

Catégorie	*Mots*	*Concepts*	*Total Paires Mots-Sens*
Nom	109195	75804	134716
Verbe	11088	13214	24169
Adjectif	21460	18576	31184
Adverbe	4607	3629	5748
Total	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 audictionnaire relationnel WordNet dans des formats multiples, aussi bien que ladécouverte des relations hiérarchiques et de traitement morphologique. Elle estcompatible avec des versions WordNet 2.0 à 3.0 et est une mise en œuvre Javacomplète.

IV.5. Corpus utilisé

Dans notre travail, nous avons utilisé le corpus Reuters10 qui est un corpus multi-label contenant 10 catégoriespossibles basé sur les 10 plus grandes catégories de Reuters-21578. Ce dernierestun ensemble de dépêches financières émises au cours de l'année1987 parl'agenceReuters, en langue anglaise, et disponible gratuitement sur le web. Cecorpus est une mise à jour du corpus Reuters-22173 qui a permis de supprimerles documents présents deux fois, de corriger des erreurs typographiques, depréciser certains formats, et de mieux définir le découpage à considérer pourl'apprentissage etle 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
wheast	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 identique 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

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

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

- Pour la deuxième approche, qui se base sur l'étape d'enrichissement nous avons utilisé les mesures de similarité sémantiquesafin 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 unseuil 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'autre mesures de similarité sémantique, de fixer certain paramètres pour en étudier d'autres plus en profondeur ainsi que plusieurs seuils. Evidement, 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 textesmultilingues*". 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 sufix stripping*". Program, 1980.

[6] Camelia Ignat, « Représentation de textes a l'aide d'étiquettes sémantiques dans lecadrede la classification automatique », European Commission, IPSC, StrasbourgFrance, 2007.

[7] Jalam, R. and Chauchat, J.-H. « *Pourquoi les n-grammespermettent de classer des textes? Recherche de mots-clefs pertinents à l'aide desn-grammes caractéristiques* ». In Morin, A. and Sébillot, P., editors,6èmes Journées internationales d'Analyse statistique desDonnées Textuelles, St. MaloFrance. 2002.

[8] Dunning, T. « *Statistical Identification of Languages ».* Technical Report MCCS 94-273, ComputingResearch 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 pourobtenir le grade de docteur d'école nationale supérieur destélécommunications, Présentée et soutenue le 09 février2007. 2004.

[12] J.Rocchio « Relevance feedback in information retrieval », 1971.

[13] Karima ABIDI, « La catégorisation de texte Multilingue », Mémoire de magistère Ecole supérieur d'Informatique, Algérie, 2010-2011.

[14] Grzegorz DZICZKOWSKI, «*Analyse des sentiments* : *systèmeautonome d'exploration des opinions exprimées dans lescritiques 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 ofthe 32nd Annual Meeting of the Associations 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 ametric 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: Anelectronic 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 intaxonomy. In Proceedings of 14th 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 toWordNet: 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 metricfor semantic similarity in wordnet. *Proceedings of ECAI'2004, the 16th European Conference onArtificial Intelligence, Valence, Espagne*. 2004.

[28] J. Jiang et D. Conrath. Semantic similarity based on corpus statistics and lexicaltaxonomy. In Proceedings of International Conference on Research inComputational 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 donnéeslexicographique WordNet et l'expérimentation est effectuée sur un corpus extrait ducorpus 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 thetexts 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,conceptualrepresentation, texts categorization,similarity measurements,WordNet

ملخص

تصنيففيالتشابهتدابيربتقييمالمتعلقالعاماالمشكلعنتتحدثالمذكرةهذه
النص.الشكليةنفسبإستخداماالفئاتوالوثائقجميعتمثيلهوالغرض.ذلكمنهدفنا
عنسابقاالمصنفةالنصوصمنمجموعةإلىإستناداامصنفةغيروثيقةكلربطهو
التشابهتدابيربطريق.البياناتقاعدةمنالنظريةالمفاهيمعلىيستندالتمثيلهذا
المعجميةwordNetمنظمة وكالة الاخبار العالميةمنمستخرجةوثائقيةقاعدةعلىالتجربةتنفيذيتمو
Reuter top 10

المفتاح:الكلمات,تدابيرتصنيف النصوص,المفاهيمتمثيل
التشابه,WordNet