République Algérienne Démocratique et Populaire

Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

UNIVERSITÉ ABOU BEKR BELKAID - TLEMCEN

FACULTÉ DES SCIENCES

DÉPARTEMENT D'INFORMATIQUE

MÉMOIRE DE FIN D'ETUDES

POUR L'OBTENTION DU DIPLOME DE MASTER EN INFORMATIQUE

Option : *Systèmes d'informations et de connaissances (S.I.C)*

Repondération des concepts en utilisant les distances sémantiques dans le cadre de la catégorisation des textes

Soutenu le 13 septembre 2017 devant le jury:

Président:	Mr Benammar Abdeklkrim	UABB Tlemcen
Examinateur:	Benmouna Youcef	UABB Tlemcen
Encadreur :	Mr Bentaallah M.Amine	UABB Tlemcen

Présenté par: **AIT ALI ALI**

SACI MOUSSA

Année académique: 2016-2017

« Certes, il y'a des travaux pénibles ;

mais la joie de la réussite n'a-t-elle pas à

compenser nos douleurs ? »

Jean de la bruyère.

Remerciements

Avant tout nous remercions DIEU de nous avoir aidé à accomplir ce modeste projet.

Nous remercions très sincèrement notre encadreur Monsieur A. BENTAALLAH pour ses remarques pertinentes et Son disponibilité permanente .

En outre ses qualités d'encadreur, nous ont toujours permis d'avancer à un rythme régulier dans notre travail et de nous encourager à persévérer.

Nous sommes très reconnaissants envers Monsieur A. BENNAMAR chef de département d'informatique.

Nous exprimons toute notre gratitude à nos examinateurs pour avoir accepter d'examiner ce travail et leurs participations au jury.

Enfin, nous ne saurions terminer ces remerciements sans y associer toute personne qui, de prés ou de loin, nous a apporté son aide ou sa sympathie.

Dédicaces

Je dédie ce travail :

A mes très chers parents qui ont toujours répondu présents dans les moments les

plus difficiles pour leurs irremplaçables et inconditionnels soutiens, leurs

confiances et leurs sacrifices qui ont contribués à ma réussite.

A mon très cher fils Mehdi ainsi mon frère Kamel et ma frangine wafaa.

A toute la famille Ait Ali.

A mon binôme Saci Moussa et toute sa famille.

A mes collègues de la section Benhmed, Mouloud, El Oulmi et je les souhaite la réussite dans leurs carrières professionnelles.

Ait Ali

Dédicaces

À la mémoire de mon défunt père.

À la plus belle créature que Dieu a créée sur terre,,,

À cet source de tendresse, de patience et de générosité,,,

À ma mère !

À tous mes frères et sœurs, ainsi que leurs enfants

À toute ma famille, mes amis et collègues Benahmed, Mouloud, El Oulmi

À tous les étudiants de la promotion 2016/2017

Option : Systèmes d'informations et de connaissance (S.I.C)

À tous ceux qui, par un mot, m'ont donné la force de continuer .....

SACI Moussa.

Table des matières

Liste des figures

Liste des tableaux

Introduction générale

L’objectif de la classification des documents est de rassembler les documents similaires qui se rapprochent des thèmes dans un même endroit afin de pouvoir par la suite de procéder à une recherche rapide et efficace.

Vu le nombre important des documents sur le web et même dans les entreprises, il est devenu nécessaire de faire une classification automatique, cette classification consiste à assigner un document a une classe avec un taux de réussite élevé et sans l’assigner à beaucoup de classes.

Il existe plusieurs méthodes de représentation des textes où chacune d’elles possède des avantages et des inconvénients, parmi ces méthodes qui présentent l’objet de notre travail est la représentation conceptuelle qui consiste à proposer une repondération des concepts.

Ce mémoire se compose de trois grandes parties, la première permet de traiter la catégorisation des textes avec toutes ses caractéristiques ensuite la seconde permet d’éclairer les mesures de similarité et enfin la dernière partie est réservée à notre travail tout en montrant les outils utilisés et les résultats obtenus avec les explications.

Chapitre 1 : catégorisation des textes

2.1 Introduction:

Notre air se distingue de la quantité impressionnante des informations sur le web, et ce le nombre de documents qui ne cesse de croitre de jour en jour, il est devenu très difficile de classer les documents avec des méthodes manuelles d'où intervient la catégorisation des textes.

Au début de ce chapitre nous allons définir la catégorisation des textes**(C.T)** puis nous allons citer ses différents types pour en finir avec une présentation des méthodes les plus connus.

2.2 Définitions

- Dans la littérature il existe différentes définitions de la catégorisation, la plus importante est celle de **[Sebastiani, 2002]** qui consiste à définir la catégorisation comme étant la tâche d'assigner une valeur booléenne pour chaque paire (d_j, c_i) Є D × C, où D est l'ensemble des documents et C est l'ensemble prédéfinie des catégories. La valeur T : « true » assignée à la paire (d_j, c_i) indique la décision de classer d_j sous c_i, tandis que la valeur F(false) indique que le document n'appartient pas à la classe C .

  Selon **[Sebastiani, 2002]** L'objectif est de construire une fonction d'assignation Φ: D × C $\rightarrow${T, F} de sorte que la décision donnée par cette fonction coïncide le plus possible avec la vraie fonction d'assignation Φ réalisée par un expert de la langue.

- Une autre définition plus récente empruntée aux sciences cognitives ciblant l'apprentissage supervisé est celle de **[Rossi, 2005, p.129]** :« Catégoriser, c'est constituer des classes d'équivalences, c'est être capable d'extraire des invariants tout en négligeant des caractéristiques non pertinentes ».

2.3 Types de la catégorisation

2.3.1 Multi/mono catégorie

Une catégorisation peut être mono ou multi label, selon **[sebastiani,2002]** la catégorisation mono-label consiste à assigner pour chaque document $d_j \in D$ exactement une seule catégorie.

Tandis que La catégorisation multi-label consiste à assigner un nombre de 0 à|C| catégories.

L'algorithme du mono-label est plus général que le multi-label il suffit juste de transformer le problème multi-label {c1….c|C| } en un problème mono-label{ci, $\overline{ci}$ }en utilisant l'algorithme de labelpowerset [Tsoumakas and Katakis, 2007] qui considère chaque combinaison de labels présente dans l'ensemble d'apprentissage comme une classe et apprend ensuite un classifieur multi-classes h_m ;par contre le contraire n'est pas vrai :un document d_j peut être classé dans plusieurs classes d'où la difficulté de déterminer la clase la plus pertinente, ou il peut se trouver que ce document n'appartient à aucune catégorie.

2.3.2 Catégorisation centrée-document, catégorisation centrée-catégorie

La C.T possède deux paramètres fondamentaux: le document et sa classe, par conséquence le processus de la C.T se base soit sur :

- le document : il s'agit de trouver toutes les classes Ci d'appartenance pour le document. dans ce cas la C.T est centrée document.
- la classe: là il s'agit de retrouver tous les documents d_j appartenant à la classe c_i.

2.4 Les étapes de la catégorisation

La catégorisation passe par plusieurs étapes formant le processus de la catégorisation (voir figure n°1) .Ce processus reçoit en entrée un texte non étiqueté et le transforme en un texte étiqueté en sortie. Afin de trouver la classe associée au texte il est nécessaire de passer par les trois phases principales **[S.L. Ting, W.H. Ip, Albert H.C.Tsang ,2011]** à savoir :

- Indexation
- Sélection des termes
- Choix du classificateur
- Evaluation du model

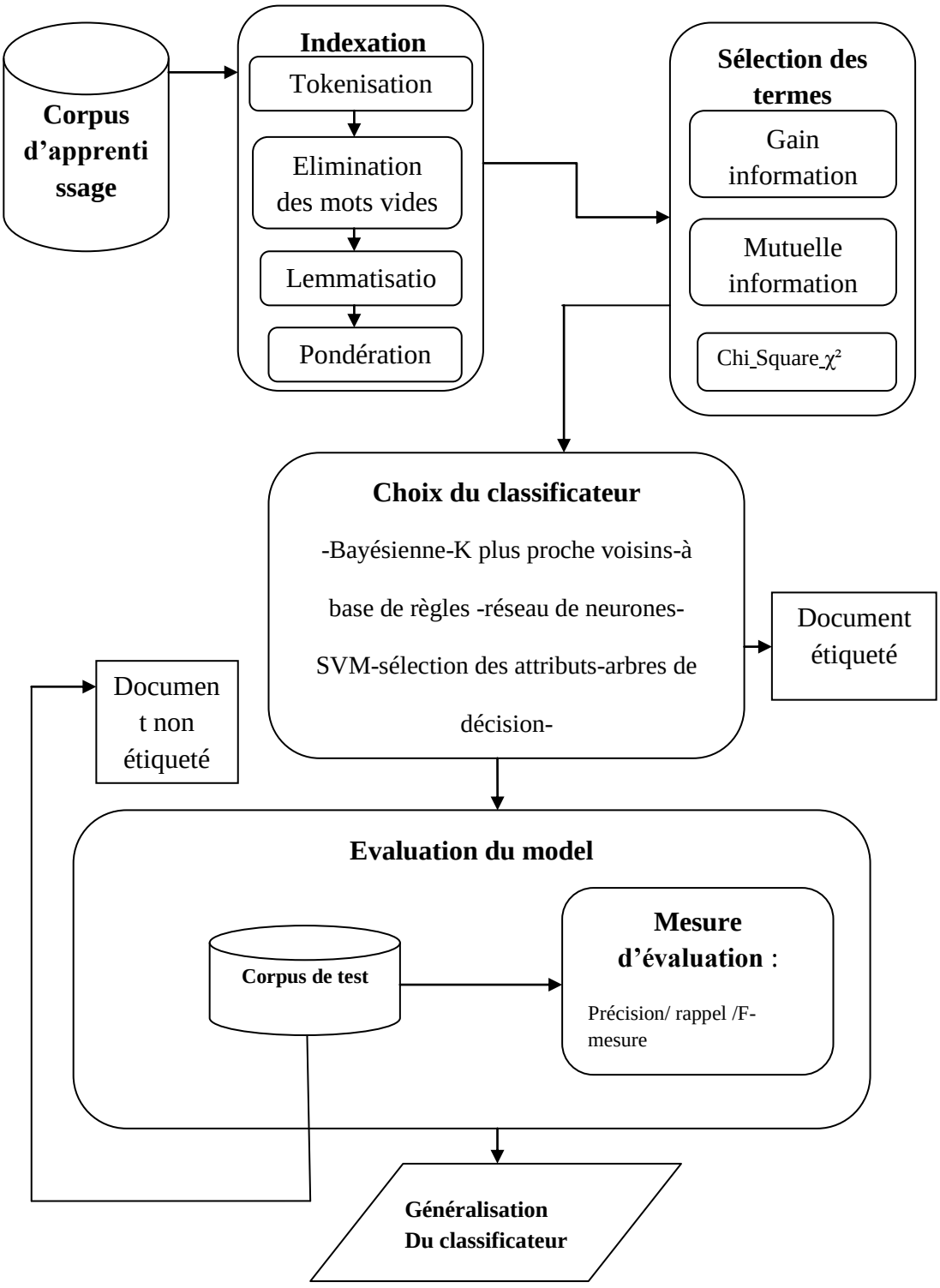


Figure N°01: processus de la catégorisation [S.L. Ting, W.H. Ip, Albert H.C.Tsang ,2011]

2.4.1 Indexation

La première étape consiste à extraire les unités représentatives du sens du corpus, dans la littérature la représentation la plus utilisée est l'indexation vectorielle qui consiste à représenter le corpus sous forme d'une matrice dont les lignes représentent les documents $d_i$et les colonnes les descripteurs t_j.Le contenu de la matrice représente le poids du terme $t_j$dans le document d_i

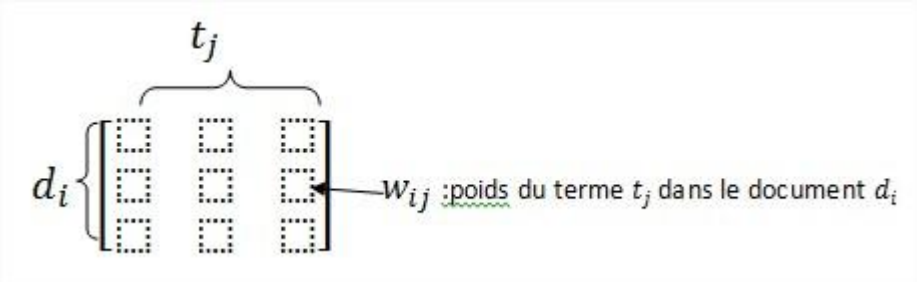


Figure N°2-matrice documents/termes

Afin de pouvoir construire cette matrice il est nécessaire de passer par les étapes suivante :

2.4.1.1 Tokenisation

Cette étape consiste à décomposer le corpus en un ensemble de descripteurs tel que des mots, des phrases ou des n-grams .La tokenisation consiste à utiliser les séparateurs de la langue (espace, majuscule, point…)pour pouvoir déterminer les descripteurs.

Ce processus de tokenisation soulève plusieurs problèmes, en effet ils existent des cas particuliers pour chaque langue où les séparations des termes fait perdre le sens du mot, exemple :Aujourd'hui, pomme de terre, les adresse IP séparés par des points(192.168.0.1), M.Mokhtar…, ainsi que certaines langues rend la tokenisation plus compliquée tel que l'allemand où une phrase n'est pas séparée par les séparateurs classiques, exemple : *Levensversicherungsgesellschaftsangestellter (employé d'une société d'assurance vie),* les langues qui s'écrivent de droite à gauche comme l'arabe mais les chiffres se lisent de gauche à droite, Le résultat de la tokenisation est un vecteur constitué des descripteurs de corpus.

2.4.1.2 Elimination des mots vides

Vu le nombre important des descripteurs il est nécessaire d'éliminer les mots qui ne portent aucun sens au document tel que les prépositions, les conjonctions, les auxiliaires…

Ces mots sont appelées les mots vides .le tableau N°1 présente un exemple de mots vides pour le français :

À	Etes	Fûtes	Votre
Allô	eurêka	Hop	Desdites
Aucuns	excepté	Pourrait	Devants
Auriez	fouchtra	pouvais	Devraient
Auxdits	fûmes	Pouvoir	Dia
Aviez	Ho	Puisse	Du
Ayons	hurrah	Pécaïre	Dut
Bof	laquelle	Quelle	Made
Çà	étaient	Tant	Mienne
Certaines	Etions	Tes	Motus
Chez	Eusse	Tout	Nôtres
Comment	Eût	unièmes	Olé
Da	Fus	Veulent	Ouille

Tableau N°1-liste des mots vides français

L'élimination des mots vide rencontre plusieurs problèmes, elle est dépendante de la langue où chaque langue possède sa propre liste, autant que certain mots existant dans la liste des mots vides peuvent ne pas être des mots vides tel que vitamine a.

2.4.1.3 Lemmatisation

Cette étape consiste à unifier les termes de différentes formes mais qui possèdent le même sens en une seule forme canonique (lemme), par exemple il est plus judicieux d'éliminer les formes :marche, marchent et marches et de les remplacer par une seule forme qui est le verbe marcher. En règle générale nous transformons les verbes en infinitif et les mots en masculin singulier.

La racinisation est l'opération de trouver la partie du mot qui reste inchangée, pour le trouver il suffit de supprimer les affixes, exemple : terminaison, terminer ↔term

Il existe plusieurs méthodes pour trouver le lemme tel que l'algorithme de porter.

La racinisation est moins sensible aux fautes d'orthographes que la lemmatisation, ainsi qu'elle n'as pas besoin de contexte pour fonctionner, uniquement le terme et la langue

L'objectif de la lemmatisation est de réduire encore plus le nombre de descripteurs tout en concentrant la pertinence.

Néanmoins la lemmatisation peut affecter la sémantique du terme, exemple: portera (verbe) et porte (nom) →porter alors que les deux termes ont des significations différentes.

2.4.1.4 Pondération

La pondération est la méthode mathématique du calcul de l'importance d'un terme dans un document, selon **[Salton & colab-1975]** la fréquence d'apparition d'un terme dans un document détermine son importance. En se basant sur la fréquence plusieurs formules existent dans la littérature dont les plus réputées :

2.4.1.4.1 La mesure TF [Jalam, 2003]

Cette mesure représente le nombre d'apparition du terme dans le document d'où sa propriété d'être locale et non globale. Selon cette mesure plus un terme est fréquent plus il est important dans le document.

2.4.1.4.2 La mesure IDF [Jalam, 2003]:

Dans cette mesure on considère que les termes les plus fréquents dans le corpus ne sont pas importants, on compense alors la fréquence des termes dans les documents (tf) par leurs fréquences dans le corpus (df)

Cette mesure est calculée par la formule suivante :

idf = log (N/df) où :

Df: le nombre de documents contenant le terme.

N: le nombre total de documents de la base documentaire.

2.4.1.4.3La mesure TFIDF [Jalam, 2003] :

Une mesure plus améliorée qui combine la tf et la idf pour prendre en compte le caractère discriminant d'un terme dans un document, elle est calculée à l'aide de la formule suivante :

TFIDF(T, D) = TF(T, D) x 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

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

Selon la tfidf pour qu'un terme t soit important dans un document d il doit se répéter fréquemment dans le document d et rarement dans les autres documents.

2.4.1.4.4La TFC [Jalam, 2003]:

L'inconvénient de la mesure TFIDF est qu'elle ne tient pas compte que les documents peuvent être de différentes longueurs, d'où l'utilisation de la TFC qui permet d'employer une longueur normalisée comme illustrée dans la formule suivante :

$$TFC(t_k, d)=\frac{TF \times IDF(t_k,d)}{\sqrt{\sum[TF \times IDF(t_k,d)]^2}}$$

2.4.2 Sélection des termes:

La sélection des termes consiste à choisir parmi un ensemble d'attribut de grande taille un sous ensemble intéressant pour la classification **[Jalam, 2003]** .

La phase de la sélection des termes est indispensable pour la catégorisation car elle facilite la visualisation des données comme elle permet de réduire considérablement l'espace de stockage dans la machine en gardant que les termes importants, ainsi de gagner en temps d'apprentissage du classificateur.

Il existe deux types de sélection des termes : supervisée et non supervisée, dans le cas de la sélection supervisée on considère que les documents sont supposés pré classifiés et il s'agit de

trouver les termes les plus importants pour chaque classe, par contre la sélection non supervisée n'exige pas de connaitre au départ le nombre de classes.

Les méthodes de sélection les plus connus sont :

2.4.2.1 Gain d'information (IG information gain)

Cette méthode possède le pouvoir de discriminer un terme dans un document, elle estime le nombre de bits d'information obtenue pour la prédiction de la catégorie en sachant la présence ou l'absence d'un mot.

Mathématiquement, ce concept se traduit par la formule suivante :

$$G(t) = -\sum_{i=1}^{m} \Pr(c_i)\log \Pr(c_i) + \Pr(t) \sum_{i=1}^{m} \Pr(c_i|t)\log \Pr(c_i|t) + \Pr(\bar{t}) \sum_{i=1}^{m} \Pr(c_i|\bar{t})\log \Pr(c_i|\bar{t})$$

où

- Pr signifie la proportion des documents ayant la caractéristique entre parenthèses
- c_i signifie qu'un document fait partie de la catégorie c_i
- t signifie qu'un document possède le terme t
- $\bar{t}$ signifie qu'un document ne possède pas le terme t
- m est le nombre de catégories

2.4.2.2 Mutuel information (MI mutuel information)

Elle mesure la quantité d'information apportée par la présence d'un terme au sujet de la présence d'un autre terme, par conséquence on peut estimer le pouvoir discriminant d'un terme dans un document. La formule suivante permet de calculer cette mesure:

Au départ on classe les documents dans des sous-groupes appelés catégories à l'aide d'un classificateur puis on calcule le poids du terme dans chaque catégorie après on obtient un tableau des poids qui vont déterminer la catégorie la plus importante.

La formule suivante permet de calculer cette mesure :

- $MI(T, C_k) = \frac{N_{T,C_k}}{N} \ln \frac{NN_{T,C_k}}{N_T N_{C_K}} + \frac{N_{T,\overline{C_k}}}{N} \ln \frac{NN_{T,\overline{C_k}}}{N_T N_{\overline{C_K}}} + \frac{N_{\overline{T},C_k}}{N} \ln \frac{NN_{\overline{T},C_k}}{N_{\overline{T}} N_{C_K}} + \frac{N_{\overline{T},\overline{C_k}}}{N} \ln \frac{NN_{\overline{T},\overline{C_k}}}{N_{\overline{T}} N_{\overline{C_K}}}$
- *T le terme*
- C_k *la catégorie*
- N_{T,C_K} nombre de présence du terme dans la catégorie C_k .
- *N* est le nombre de document total.
- N_{C_K} est le nombre de document dans la catégorie C_k.
- $N_{\overline{C_k}}$ est le nombre de document hors catégorie C_k.
- N_T est le nombre de présences du terme dans tout le corpus.
- $N_{\overline{T}}$ est le nombre d'absence du terme dans tout le corpus
- $N_{\overline{T},C_K}$ est le nombre d'absence du terme dans C_k.
- $N_{T,\overline{C_K}}$ est le nombre de présences du terme hors C_k.
- $N_{\overline{T},\overline{C_K}}$ est le nombre d'absence du terme hors C_k.

2.4.2.3 ***Chi Square χ²***

Cette méthode est basée sur le test des couples termes/catégories.

Il existe d'autres méthodes tel que : corrélation coefficient, odds ration, DIA facteur d'association, relevancy score, GSS coefficient.

Cette méthode est calculée par la formule suivante :

$$\chi^2(T, C_k) = \frac{N \times \left(\left(N_{T,C_k} \times N_{\overline{T},\overline{C_k}} \right) - \left(N_{T,\overline{C_k}} \times N_{\overline{T},C_k} \right) \right)^2}{N_T \times N_T \times N_{C_k} \times N_{\overline{C_k}}}$$

2.4.3 Choix du classificateur :

Après les phases précédentes le nombre des termes restants est considérablement réduit et sont très précis pour la construction d'un modèle de classification.

Dans l'étude de **[S.L. Ting, W.H. Ip, Albert H.C.Tsang]** ils ont choisi le classificateur Bayésien Naïve à cause de sa simplicité ainsi sa performance.

Il existe plusieurs techniques dont les plus connus sont les suivants **[Ph. PREUX, 2011]:** Arbres de décision, bayésiennes, k plus proches voisins, réseaux de neurones, SVM…

2.4.3.1 Classification par arbres de décision

[LAHLOU OUCHIHA, 2016] comme son nom l'indique c'est un ensemble de nœuds reliés par des branches, la racine se trouve en haut et les feuilles en bas. L'arbre de décision est multi-classes, le premier descripteur placé dans la racine est le plus discriminant des textes du corpus. L'opération est répétée pour les nœuds suivants jusqu'à ce qu'on peut plus séparer les textes et on obtient des feuilles qui contiennent des textes de la même classe, comme illustré dans la figure N°3.

Dans cette méthode on considère un ensemble X d'exemples contenant différents attributs et on veut les classer dans une classe(ou attribut cible), on construit alors un arbre de décision qui se compose de :

- Nœuds: représentent les tests sur les attributs
- Les branches: représentent les valeurs de ces tests.
- Les feuilles: représentent les valeurs des attributs cibles

Dans cet exemple en prenant un ensemble de jours(1jour= exemple)chaque exemple se caractérise par l'état de la météo(température, vent, ciel, humidité)l'attribut cible est

donc : « jouer au tennis » qui peut prendre que deux valeurs « oui » ou « non » et nous parlons alors d'une classification binaire.

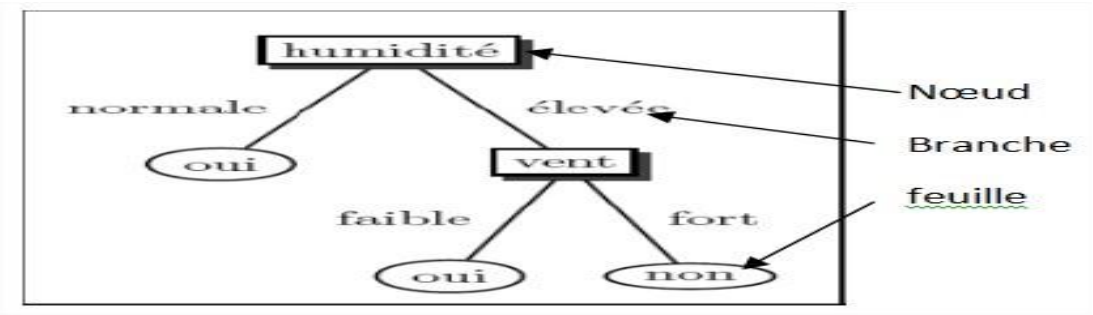


Figure N°3 : exemple d'arbre de décision jouer au tennis

Parmi les avantages de cette méthode est qu'elle ne nécessite pas d'hypothèse sur les données, et les résultats sont plus simples à exploiter, elle est transparente pour l'utilisateur et son temps d'apprentissage est relativement court.

Par contre son inconvénient est qu'elle fournit souvent des arbres instables (une division conditionne les suivantes, les branches coupées ne repoussent pas), ainsi dans le cas des arbres volumineux elle devient peu explicatif et elle est sujette du sur apprentissage (apprentissage par cœur).

2.4.3.2 Classification bayésienne

Selon **[Ph. PREUX, 2011]** à la différences des autres méthodes dites fréquentielles, cette méthode est probabiliste, elle estime Pr[x|w] :la probabilité de l'occurrence qu'un évènement x si l'évènement w est vérifié, par exemple dans le monde réel, on peut calculer la probabilité d'agir d'une façon négative(évènement x) sachant qu'on est dans la situation w.

Selon Bayes, la probabilité de A, sachant B et C où C représente la connaissance que l'on a :

Où :

$$Pr[A|B,C] = \frac{Pr[B|A,C]Pr[A|C]}{Pr[B|C]}$$

- Pr[A|B,C] :la probabilité à posteriori de A si B et C sont vérifiés
- Pr[B|A,C] :la probabilité de B si A et C sont vérifiés
- Pr[A|C] :la probabilité a postériori de A si C est vérifié
- Pr[B|C] :la probabilité marginale de B si C est vérifié

En appliquant le théorème de Bayes à la catégorisation des textes on obtient :

Pr[y|x,X] :la probabilité d'appartenance à une classe y en observant la donnée x tout en disposant d'un ensemble d'exemples X est calculée comme suit :

$$Pr[y|x,\mathcal{X}] = \frac{Pr[x|y,\mathcal{X}]Pr[y|\mathcal{X}]}{Pr[x|\mathcal{X}]}$$

Où :

Pr[y|X] :estimer la probabilité d'appartenance de l'ensemble d'exemples X à une classe y

Pr[x|y,X] :estimer l'appartenance d'une donnée x à une classe y tout en disposant d'exemples X, et pour faciliter le calcul nous utilisons l'hypothèse de bayésien naïve qui suppose que la donnée x composée de plusieurs attributs aléatoires et indépendants les uns des autres.

Pr[x,X] :la probabilité d'observer la donnée x en disposant de l'ensemble X.

Parmi les avantages de Bayse **[I.Telier-Paris3]** est la rapidité, simple et efficace pour les données textuelles, il est facile de faire des mises à jour pour classer de nouveaux documents et par conséquence il est très utilisé dans la gestion des courriers électroniques (spam ou non spam).

Par contre ce classificateur est difficilement lisible par l'humain à cause de son calcul statistique, ainsi que ces calculs n'ont pas de sens que si on fait une hypothèse ce qui est exagérée.

2.4.3.3 Classification à base d'exemples présentatifs(K plus proches voisins)

Il s'agit dans cette méthode de trouver les cas exemples qui se rapprochent le plus de l'élément à classer **[Ph. PREUX, 2011]**. En effet l'élément sera classé dans la classe la plus représentative parmi les cas exemples les plus proches. Afin de pouvoir calculer le rapprochement entre les exemples plusieurs distance existent tel que :

- Distance euclidienne
- Distance Manhatan
- Distance Maximale

L'algorithme du k plus proches voisins est énoncé comme suit :

pour chaque nouveau point x on commence par déterminer l'ensemble de ses k-plus proches voisins parmi les points d'apprentissage que l'on note Vk(x)(on doit choisir $1 \leq k \leq n$).

La classe que l'on affecte au nouveau point x est alors la classe majoritaire dans l'ensemble Vk(x).

Cette méthode est réputée dans le domaine commercial utilisant les systèmes de recommandation qui permettent de prédire le comportement des clients en présence de certains produits.

Parmi les inconvénients de cette méthode est le choix du voisinage, en effet il est difficile de déterminer le nombre de voisins à retenir, ainsi que le choix de la meilleure distance ce qui n'est pas évident.

2.4.3.4 Classification à base de règles

Cette méthode a pour objectif de construire un classeur composé de règles de classification, pour cela il est nécessaire d'utiliser un ensemble de règles ordonnées R1, R2… de type :

Si condition(x) alors classe(x) = C

Pour classer une donnée on vérifie si la donnée vérifie la première règle sinon on vérifie la deuxième règle et ainsi de suite **[Ph. PREUX, 2011]**.

Plusieurs approches ont étaient proposées, les plus connus sont :C4.5 rules, Prism, règles d'association.

Dans cette approche il s'agit de construire des ensembles d'items attribut=valeur (exemple : température=chaud). A chaque item sera associé une fréquence représentant le de nombre d'occurrence de cet item dans le jeu d'exemple, la classe est considérée comme un item.

Afin de pouvoir construire les règles il s'agit de fixer une fréquence S qui représente le nombre d'objets auxquels la règle s'applique permettant de construire des ensembles selon la valeur de l'item(item>s).

Supposant que H I J K L cinq items qui vérifient {H,I,J} {H,I,K} {H,J,K} {H,J,L} {I, J, K}>S alors il est fort possible que {H,I,J, K} ≥S

L'étape suivante consiste à tirer les règles d'associations comme suit :

si H et I alors J
si H et J alors I
si I et J alors H
si H alors I et J
si I alors H et J

si J alors H et I
si Ø alors H, I et J

L'étape suivante est de calculer la précision pour chaque règle

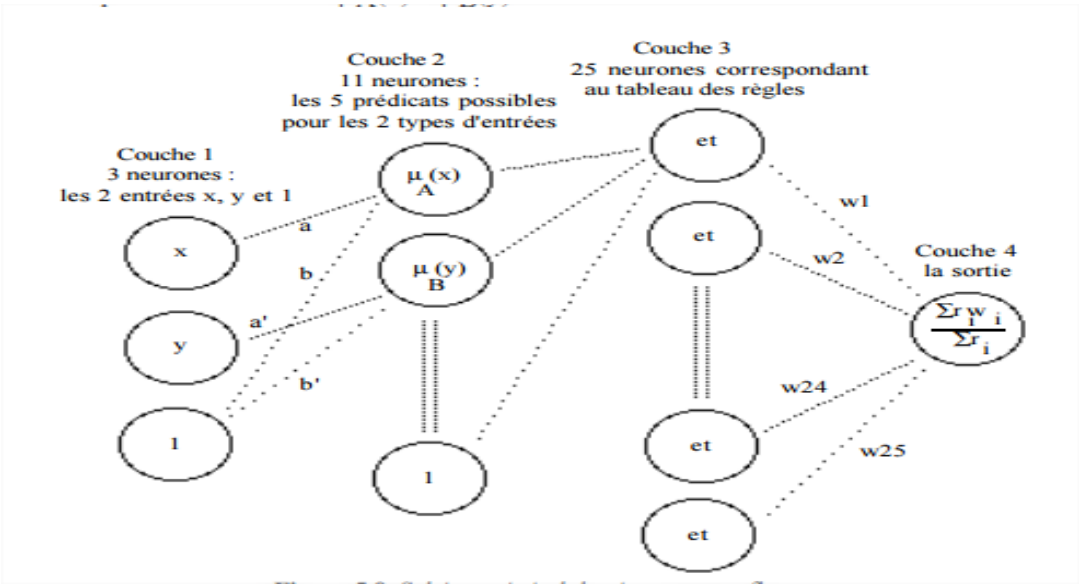
$$P = \frac{\text{nombre d'exemples couverts pour lesquels la conclusion de la règle est correcte}}{\text{nombre d'exemples couverts par la règle}}$$

L'avantage des algorithmes de règles d'association est le fait qu'il puisse exister des associations entre tous les attributs.

Par contre leurs inconvénient est qu'ils recherchent des éléments dans un espace de recherche très grand, nécessitant ainsi beaucoup plus de temps.

2.4.3.5 Classification par réseaux de neurones

[Ph. PREUX, 2011] il existe plusieurs types de réseaux de neurones tel que perceptron (neurone formel), Kohonen(apprentissage non supervisé) et cela est illustré sur la figure N°4.



FigureN°4-reseaux de neurones

Le perceptron est une unité de traitement composée de :

- P+1 entrées notée $e_i \in \{0,..,P\}$
- une sortie s
- un poids Ш pour chaque entrée $Ш_i \in \{0,..,P\}$

- une fonction $s = \varphi(\sum_{i=0}^{i=P} w_i e_i)$ d'activation qui détermine la valeur de la sortie s en fonction des entrées et leurs poids
- le potentiel v= $\sum_{i=0}^{i=P}$
- le perceptron prédit la classe en sortie s en fonction des données qui sont les entrées e_i.

Dans le travail de **[Mathieu Stricker,2004]** il existe trois types de la fonction d'activation :

La fonction identité : f(v) = v.

La fonction sigmoïde : f(v) = tanh(v) « tangente hyperbolique ». C'est une fonction bornée à valeurs réelles comprises entre -1 et +1.

La fonction logistique : f(v) = 1/(1 + exp(-v)). C'est une fonction bornée à valeurs réelles comprises entre 0 et 1.

Dans le cas des problèmes multi classes on utilise une combinaison de plusieurs réseaux de neurones formels interconnectés pour obtenir plusieurs sorties **[Mathieu Stricker,2004]**.

Pour l'apprentissage du réseau de neurone il s'agit de déterminer les poids pour avoir une sortie s la plus proche de l'objectif initial, dans le cas d'une classification supervisée nous somme réduit à déterminer une surface de séparation.

L'apprentissage est réalisé grâce à la fonction de coût qui utilise une base d'apprentissage et la sortie du réseau de neurone.

2.4.3.6 Classification par SVM(supports vector machine)

Selon ce classificateur il s'agit trouver un séparateur qui divise le corpus en deux catégories et que la largeur de ce séparateur (marge) soit maximale **[LAHLOU OUCHIHA, 2016]**.

La figure N°5 illustre le travail du MVS

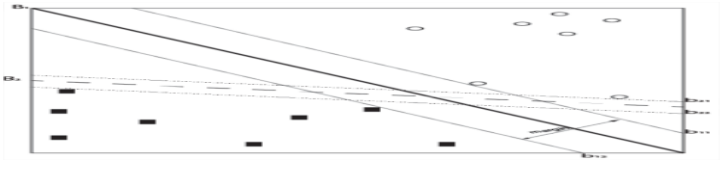


Figure N°5-classificateur SVM

Les vecteurs de support sont les points les plus proches de la marge.

Les SVM sont utilisés dans différents domaines tel que la classification de données biologiques/physiques, la classification d'expressions faciales, la classification de textures, le E-learning, la détection d'intrusion et la reconnaissance de la parole.

Parmi les limites des SVM est qu'ils sont bi-classes et dans le cas d'un problème multi-classe nous somme obligé de le transformer en bi-classes. Le temps d'apprentissage est relativement lent par rapport aux arbres de décisions.

2.4.3.7 Classification par sélection des attributs

La sélection des attributs permet de construire un sous ensemble des attributs intéressants parmi un grand ensemble **[Ph. PREUX, 2011]**.

Dans cette méthode on considère que la classification est binaire et les attributs sont de type quantitatif.

La fonction a(x) détermine la valeur de l'attribut a de la donnée x, pour couvrir tous les attributs on utilise la fonction a_j où a_j est la fonction renvoyant le jemme attribut a

On cherche un classeur définie par la fonction $\hat{y}(x) = \sum_{a \in \mathcal{A}} w_a a(x).$

Où w_a représente le poids de l'attribut a.

Le signe de la fonction détermine la classe prédite pour la donnée x.

Etant donné les valeurs des attributs qui sont fixes il reste à déterminer les valeurs des w_a .

2.4.4 Evaluation du model:

Une fois le model construit dans l'étape précédente il est nécessaire d'évaluer le résultat de la classification. Afin de pouvoir évaluer une C.T, plusieurs mesures existent dans les littératures, la majorité de ces mesures se basent sur deux notions importantes à savoir la précision et le rappel.

La précision consiste à mesurer la proportion des documents pertinents par rapport à l'ensemble des documents renvoyés par le système, tandis que le rappel mesure la proportion de documents pertinents retournés par le système par rapport à l'ensemble des documents pertinents se trouvant dans la base documentaire, et voici les deux formules :

$$\text{Précision}=\frac{|Ra|}{|A|}$$

$$\text{Rappel}=\frac{|Ra|}{|R|}$$

Où : |A| : le nombre de documents retournés par le système pour une requête donnée

|R| : le nombre de documents pertinents dans une collection pour cette requête

|Ra| : le nombre de documents pertinents retournés par le système

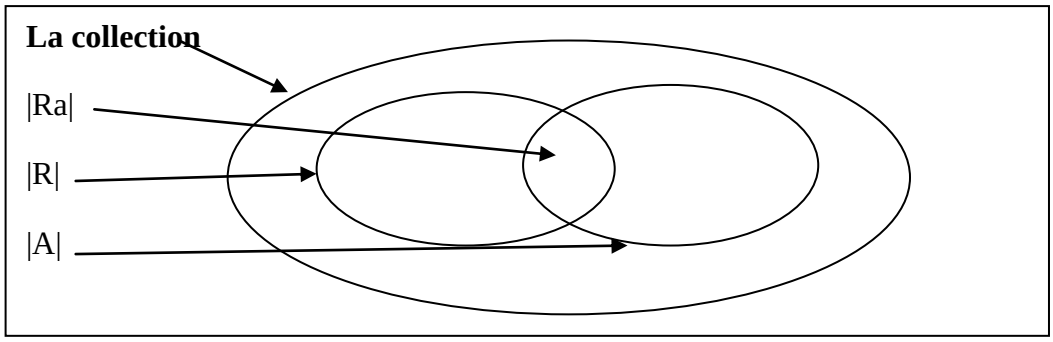


Figure N°06-Rappel/Précision

Lorsqu'un système retourne un résultat par rapport à un document et une classe deux possibilités sont présentes :

- Le document appartient selon le système à la classe.
- Le document n'appartient pas à la classe selon le système.

Par contre en réalité on a les deux possibilités suivantes :

- Le document appartient à la classe.
- Le document n'appartient pas à la classe.

On peut déduire alors quatre possibilités illustrées dans le tableau suivant :

Nom du cas	Abréviation	Description
Vrai positif	VP	Le système trouve à raison le document comme appartenant à la classe
Faux positif	FP	Le système trouve à tort le document comme appartenant à la classe
Vrai négatif	VN	Le système trouve à raison le document comme n'appartenant pas à la classe
Faux négatif	FN	Le système trouve à tort le document comme n'appartenant pas à la classe

Tableau N°2-cas des possibilités des doc retournés

De ce principe on peut réécrire la formule précision/rappel en fonction de ces paramètres comme suit :

Précision=$\frac{vp}{vp+fp}$

Rappel=$\frac{vp}{vp+fn}$

Une autre mesure qui permet de combiner rappel/précision en un nombre compris entre 0 et 1 est la moyenne harmonique F-mesure donnée par la formule suivante :

F-mesure=$\frac{2PR}{P+R}$

On peut calculer la similitude et la diversité entre deux populations à l'aide de la formule de Jaccard formulée comme suit :

Jaccard=$\frac{vp}{vp+fp+fn}$

La moyenne harmonique peut être pondérée pour favoriser sois le rappel ou la précision par la E-mesure donnée par la formule suivante :

$F_\beta=\frac{(1+\beta^2)P.R}{\beta^2.P+R}$

Si β>1 on privilège la précision

Si β<1 on privilège le rappel

Si β=1 on est ramené à calculer la F-mesure

Dans le cas où le résultat n'est pas satisfaisant on doit modifier la technique de la sélection des termes ou nous remettons en question le choix du classificateur (Bayésien Naïve).

Dans le cas de plusieurs requêtes pour le système on peut calculer la moyenne de deux façons :

- Micro-moyenne: consiste à faire la somme des vrais positifs, faux positifs et faux négatifs avant de calculer les valeurs de rappel et de précision

  Micro-rappel = $\frac{\sum_{i=1}^{n} vp}{\sum_{i=1}^{n} vp + \sum_{i=1}^{n} fn}$

  Micro-précision = $\frac{\sum_{i=1}^{n} vp}{\sum_{i=1}^{n} vp + \sum_{i=1}^{n} fp}$

- Macro-moyenne: Consiste d'abord à calculer les valeurs de rappel et de précision sur chacune des n classes avant d'en effectuer une moyenne

  Micro-rappel = $\frac{\sum_{i=1}^{n} [\frac{vp(i)}{vp(i)+fn(i)}]}{n}$

  Micro-précision = $\frac{\sum_{i=1}^{n} [\frac{vp(i)}{vp(i)+fp(i)}]}{n}$

2.5 Conclusion

Dans ce chapitre on a essayé de dresser un panorama des classificateurs afin de pouvoir sélectionner le plus efficace pour la partie de l'expérimentation.

Le processus de catégorisation est devisé en plusieurs étapes où chaque étape participe à la réduction des termes les plus importants pour concentrer la pertinence.

Donc l'évaluation dépend de plusieurs paramètres c'est pour cette raison qu'il est difficile d'évaluer, contrairement à la catégorisation où le problème est localisé soit dans une étape ou une autre.

Chapitre 2 : Mesures de similarité

3.1 Introduction

Les mesures de similarité statistique reposent essentiellement sur des valeurs où il n'existe pas de dépendance entre les mots, alors que dans la réalité on peut trouver des documents très proches mais qui n'utilisent pas les mêmes mots (synonymes) par conséquence l'utilisation de ces mesures n'assure pas une bonne similarité, en effet la notion sémantique est souvent ignorée dans ce type de mesure **[Elsa Negre-2013]** comme par exemple il n'y a pas de similarité entre « j'ai un chien » et « je possède un animal », ainsi que la relation syntaxique n'est pas prise en compte, dans l'exemple suivant les deux textes sont considérés similaires alors qu'ils sont en réalité différents: « Christine aime Simon » et « Simon aime Christine ». De plus les problèmes de la négation qui conduit à une similarité élevée alors que c'est totalement le contraire, exemple : « je suis à la maison » et « je ne suis pas à la maison ».

Les mesures de similarité sémantique permettent de palier les problèmes précédemment cités. EN effet, il existe plusieurs travaux sur les mesures de similarité sémantique dans le but d'améliorer et faciliter l'accès à l'information de façon pertinente. D'après [Thabet et al,2007] on peut identifier trois grandes familles d'approches pour l'identification de la similarité sémantique. Les approches basées sur les nœuds [Res, 1995][Lin, 1998][Jiac, 1997] définissant la similarité conceptuelle en s'appuyant sur des mesures du contenue informationnel. En plus, le degré de partage de l'information détermine la similarité entre les concepts. La deuxième famille d'approche s'appuie uniquement sur la hiérarchie ou sur les distances des arcs [Rada, 1989][Lee, 1993] [Wup, 1994]. Le calcul de la similarité dans cette approche est basé principalement sur la longueur du chemin entre les nœuds, plus le chemin est court plus les nœuds sont semblables. La deuxième caractéristique de cette approche est que les arcs d'une taxonomie ont une longueur uniforme, ce qui implique l'inconvénient que tous les liens sémantiques possèdent le même poids ce qui impose des difficultés au niveau de la définition et du contrôle des distances des liens [Thabet et al, 2007]. La troisième est l'approche hybride [Lec 98][Res99] qui combine entre les deux approches cités en-dessus,

cette approche présente une manière différente pour la détermination de la similarité conceptuelle entre deux mots dans un réseaux sémantique hiérarchique.

3.2 Définition:

Selon [Zargayouna, 2005] la similarité entre deux concepts dans une ontologie représente une fonction inverse de la distance entre eux, ce qui implique que plus deux concept sont distants moins ils sont similaires et vise vers ça. La distance entre les concepts représente une mesure qui doit respecter les trois caractéristiques suivantes:

- la distance d'un concept avec lui-même doit être nulle: dist (x,x) =0.

- symétrie: dist(x,y) = dist(y,x).

- inégalité triangulaire: dist(x,y) ≤ dist(x,z) + dist(z,y).

[Zargayouna,2005] partage les mêmes pensés que [lin, 1998] face aux caractéristiques que devrait respecter la notion de similarité, donc la similarité entre deux concepts X et Y doit être:

- fonction de la caractéristique commune. Plus le nombre des caractéristiques communes augmente plus les concepts sont similaires.

- fonction de leurs différences, plus le nombre des caractéristiques différentes augmente plus la similarité diminue.

- maximale si X et identique de B.

" Une mesure de similarité est une fonction Sim: $S^2 \rightarrow [0.1]$ avec S l'ensemble des concepts." [Zargayouna,2005].

Les points de vue divergent face aux propriétés que doivent respecter les mesures de similarité, cependant, [Tversky, 1977] a prouvé que ces propriétés ne sont pas toujours

satisfaites par les mesures de similarité conformes à la perception humaine, par exemple la symétrie. Donc en générale il est admis que les mesures de similarité doivent être réflexives et symétriques:

- Sim (x,x) = 1. réflexivité.

- Sim (x,y) = Sim (x,y) symétrie.

3.3 CLASSIFICATION DES APPROCHES DE MESURES DE SIMILARITE:

Les approches des mesures de similarité sont classifiées par [Thabet et al, 2007] selon le schéma ci-dessous:

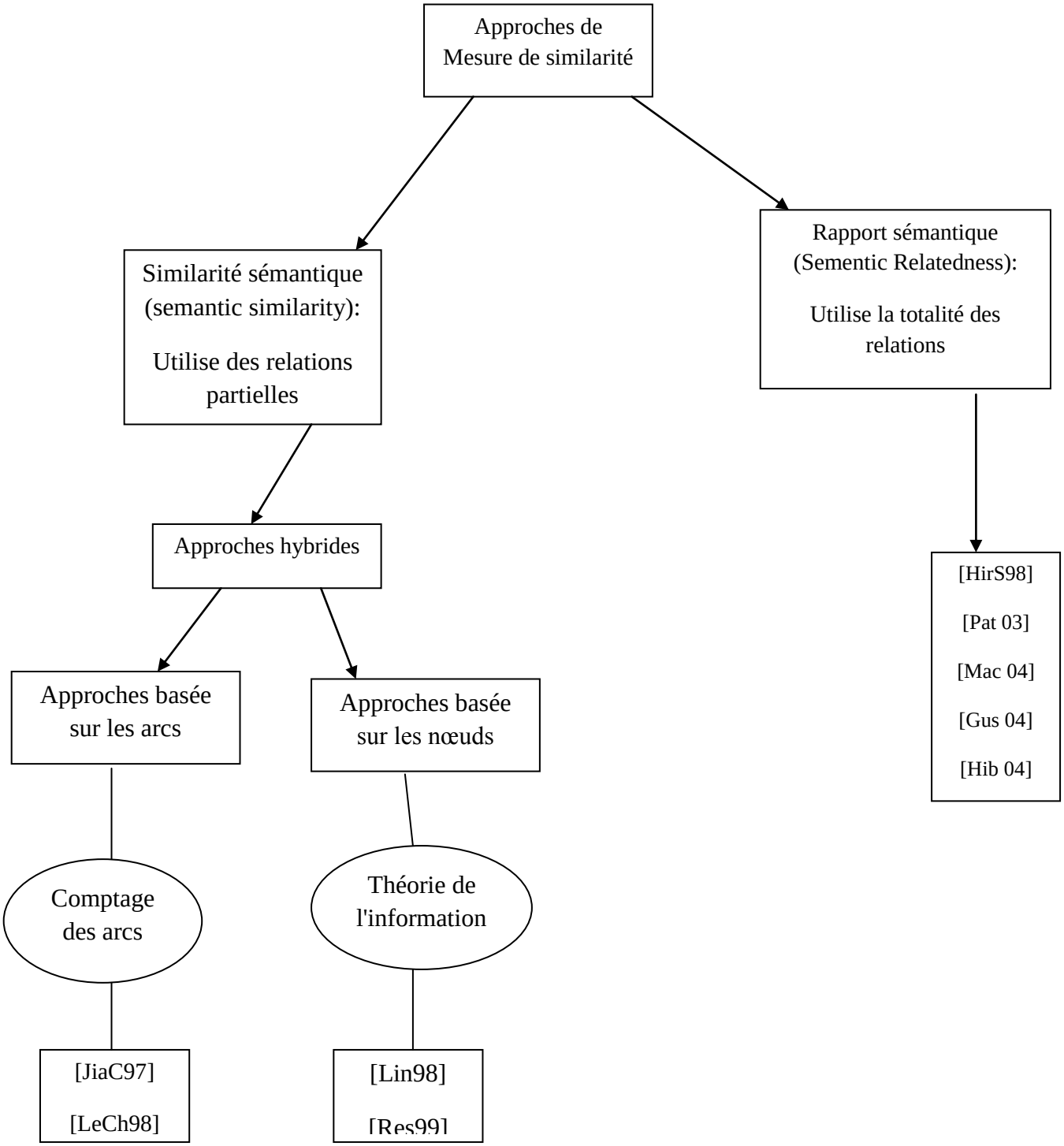


Figure 07- *Taxonomie des approches de mesure de similarité.*

3.3.1 Approches basées sur les arcs:

Selon [Andon,2012] les mesures de similarité basées sur les arcs ou en peut dire aussi basée sur la distance taxonomique définie la similarité entre deux sens d'une taxonomie par le comptage du nombre d'arcs qui les séparent.

"Ces mesures se servent de la structure hiérarchique de l'ontologie pour déterminer la similarité sémantique entre les concepts"[Thabet et al,2007]. En effet, le calcul des distances dans l'ontologie est basé sur un graphe de spécialisation des objets, ces distances doivent être caractérisées par le plus court chemin qui fait intervenir le plus petit généralisant c'est à dire l'ancêtre commun qui connecte deux objets à travers des descendants communs. Parmi les travaux recensés sous cette bannière on peut citer:

3.3.1.1 Mesure de Wu et Palmer:

Selon [Thabet et al,2007] le principe de la mesure de similarité de [Wup,1994] est le suivant:

Etant donnée une ontologie Ω formée par un ensemble de nœuds et un nœud racine R(figure N°8). Soit A et B deux éléments de l'ontologie que nous allons calculer la similarité entre eux, la similarité est calculée selon un principe où en se basant sur les distances (N1 et N2) qui séparent les nœuds A et B du nœud racine et la distance qui sépare le concept subsumant[1] (CS) de A et B du nœud R.

[1] Le concept subsumant c'est le concept commun le plus spécifique.

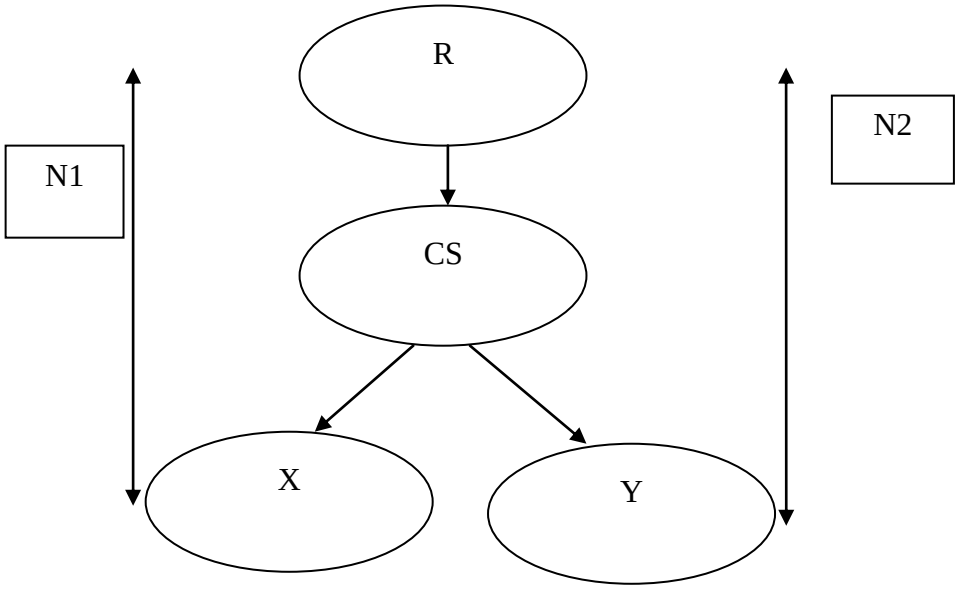


Figure -08- *Exemple d'un extrait d'ontologie.*

La mesure de Wu et Palmer est définie par la formule suivante:

$$Sim(A,B) = \frac{2 * N}{N1 + N2 + 2 * N}$$

Selon [Zargayouna et Salotti,2004], 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. Voici une formule plus formelle de cette mesure:

$$Sim(A,B) = \frac{2 * depth(CS)}{depth_{(CS)}(A) + depth_{CS}(B)}$$

Où CS est le PPG de A et B (en nombre d'arcs), depth(C) est le nombre d'arcs qui sépare CS de la racine et $depth_{CS}(A)$ et $depth_{CS}(B)$ est le nombre d'arcs qui séparent les deux concepts A et B de la racine en passant par CS.

Après une étude comparatifs entre les mesures de similarité [Lin,1998] a prouvé que la mesure de Wu & Palmer [Wup, 1994] a l'avantage d'être simple à implémenter en plus des performances qu'elle présente.

"La mesure de [Wup, 1994] est intéressante mais présente une limite car elle vise essentiellement à détecter la similarité entre deux concepts par rapport à leur distance de leur plus petit généralisant. Ce qui ne permet pas de capter les mêmes similarités que la similarité conceptuelle symbolique " [Thabet et al, 2007].

D'après [Thabet et al, 2007] et [Zargayouna et Salotti, 2004] le problème major de la mesure de [Wup, 1994] c'est qu'avec cette mesure on peut obtenir une similarité plus élevée entre un concept et son voisinage par rapport à ce même concept et un concept fils, ce qui est inadéquat dans le domaine de la recherche de l'information.

Cette limite de la mesure de [Wup, 1994] a poussé les deux chercheurs précédents de proposer d'autres formules (mesures) qui s'inspirent des avantages du travail de [Wup,1994] et qui corrigent ses inconvénients par la pénalisation des concepts de voisinage et la prise en compte des concepts fils en priorité, donc aucun concept du voisinage ne sera plus similaire que les fils.

3.3.1.2 La mesure de [Zargayouna et Salotti, 2004]:

Zargayouna et Salotti ont défini une fonction spec(C1,C2) qui calcule la spécificité de deux concepts par rapport au concept le plus bas de l'ontologie (bottom) comme le montre la figure N°9. Cette fonction servira à pénaliser les concepts qui ne sont pas dans la même ligne c'est à dire elle assure de prendre en compte les fils en priorité.

$spec(C1, C2) = N4 * N1 * N2$ (Voir figure 9).

Plus formellement:

$$spec(C1, C2) = depth_b(C) * distance(C, C1) * distance(C, C2)$$

avec $depth_c(C)$ est le nombre maximum d'arcs qui séparent C de bottom et $distance(C,C_i)$ est la distance en nombre d'arcs entre C et C_i.

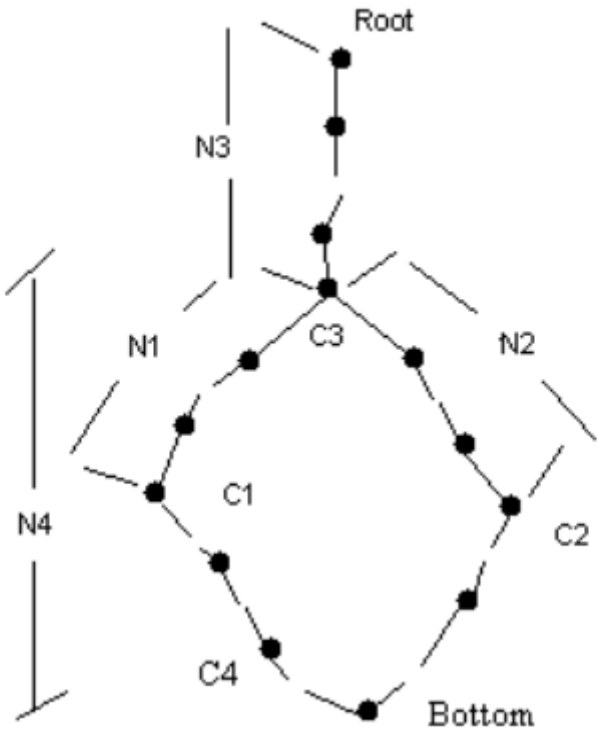


Figure 09- *Les nouvelles relations conceptuelles*

La nouvelle mesure de similarité devient:

$$sim(C1, C2) = \frac{2 * depth(C)}{depth_c(C1) + depth_c(C2) + spec(C1, C2)}$$ [Haifa et Sylvie, 2004].

3.3.1.3 <u>La mesure de [Thabet et al, 2007]:</u>

[Thabet et al, 2007] a défini une fonction de pénalisation $f_p(C1, C2)$ qui assure la pénalisation de la similarité de deux concepts éloignés qui ne sont pas dans la même hiérarchie.

$$\begin{cases} f_p\,(C1,C2) = 1 \; si \; C1 \; est \; ancêtre \; de \; C2 \; ou \; l'inverse. \\ f_p(C1,C2) = \dfrac{1}{|profondeur(C1) - profondeur(C2)| + 1} \; Autrement. \end{cases}$$

Donc la nouvelle mesure est représentée par la formule suivante:

$Sim_{tbk} = \frac{2*profondeur\,(C)}{Profondeur\;(C1)+profondeur(C2)} * f_p\,(C1,C2)$ [Thabet et al, 2007].

Où :

Profondeur (C_i) est la distance en nombre d'arcs entre C_i et le concept subsumant commun (C).

3.3.1.4 La mesure de Rada et al:

" Cette mesure [Rada et al, 1989] est adoptée dans un réseau sémantique et elle est fondée sur le fait qu'on peut calculer la similarité en se basant sur les liens hiérarchiques «is-a»" [Thabet et al, 2007].

D' après [Thabet et al, 2007] le principe de calcul de la mesure de [Rada et al, 1989] est de calculer la distance entre les nœuds par le chemin le plus court, c'est à dire pour calculer la similarité entre deux concepts dans une ontologie on doit calculer le nombre d'arcs minimums qui les séparent.

Selon [Andon, 2012] la mesure de [Rada et al, 1989] considère la similarité directement comme la distance entre les nœuds correspondants aux deux sens, mais ne considérant que les liens hyperonymie et hyponymie. Elle s'exprime par:

$Sim_{Rada}(S1,S2) = d(S1,S2) = N1 + N2$. [Rada et al, 1989].

3.3.2 Approches basées sur les nœuds:

"Ces approches adoptent une nouvelle mesure en termes de la mesure entropique de la théorie de l'information [Lin 1998] [RES 1999]."[Thabet et al, 2007].

L'information de la classe est désignée par la probabilité P(.) pour l'identification de l'utilisation d'une classe ou de ses classes filles dans un corpus. L'entropie d'une classe est définie par la formule suivante:

$$E(c) = -\log \mathrm{P}(P(c))$$

où: P représente la probabilité de trouver une instance du concept et elle est calculée en divisant le nombre des instances du concept par le nombre total des instances.

Selon [Thabet et al, 2007], il est possible d'éviter le manque de fiabilité des instances des arcs, en associant des probabilités aux concepts d'une taxonomie. Parmi les travaux classifiés sous cette bannière on peut citer:

3.3.2.1 La mesure de Resnik:

Selon [Zargayouna et Salotti, 2004] la notion du contenu informationnel (CI) a été la première fois introduite par [Resnik, 1995], elle utilise à la fois l'ontologie et le corpus.

D'après [Thabet et al, 2007], [Resnik, 1995] a prouvé qu'un objet (mot) est défini par le nombre des classes spécifiées et que la quantité d'information partagée entre deux concepts mesure la similarité sémantique entre eux, donc il est obligatoire de calculer le contenu informationnel à fin d'évaluer la pertinence d'un objet.

[Zargayouna et Salotti, 2004], a mentionné que la pertinence d'un concept dans un corpus est traduite par le contenu informationnel de ce concept en tenant compte de sa spécificité ou généralité, en plus pour trouver le contenu informationnel il faut calculer la fréquence du concept dans le corpus qui représente l'apparition du concept lui-même et l'apparition des concepts qu'il *subsume*[2]. Cette mesure est définit par la formule suivante:

$$CI(c) = -\log \mathrm{P}(P(c))$$

[2] On dit qu'un concept général subsume un concept plus spécifique.

avec P(c) est la probabilité de trouver une instance du concept c et elle est calculée selon la formule suivante:

$$\frac{fréquence(c)}{N}$$

N est le nombre de tous les concepts dans le corpus.

La figure ci-dessous illustre un extrait de WordNet où chaque nœuds est attaché à sa probabilité P(c).[Lin, 1998]

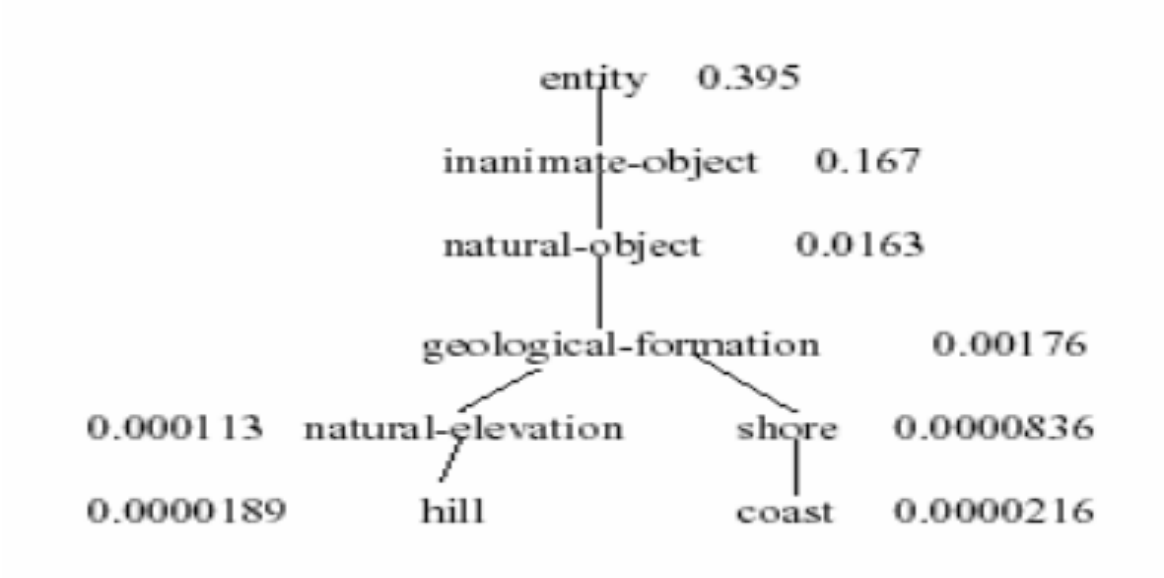


Figure-10- Extrait du WordtNet.

La quantité d'information partagée par deux concepts représente la similarité sémantique entre eux et elle est égale au contenu informationnel du plus petit généralisant (PPG)[3] selon Resnik.[Zargayouna et Salotti, 2004]. En effet, cette mesure est définie par cette formule: $Sim(c1,c2) = CI(ppg(c1,c2))$.

Selon [Haifa et Sylvie, 2004] la mesure de resnik est un peu sommaire car elle est dépendante absolument au PPG ce qui nous donne ppg(x,y)=ppg(a,b) même si a et b sont plus proches que x et y du PPG.

[3] le concept le plus spécifique qui subsume deux concepts dans l'ontologie.

3.3.2.2 La mesure de Lin:

La mesure de similarité de [Lin, 1998] est légèrement différente de celle de Resnik et elle est définie par:

$$Sim(x,y) = \frac{2 * \log(P(AC(x,y)))}{\log(P(x) + P(y))}$$

Selon [Thabet et al, 2007] cette mesure utilise deux source de connaissances différentes (Thésaurus et corpus) donc elle utilise une approche hybride, en plus le degré de probabilité de chevauchement des concepts descendants de X et Y représente cette similarité.

Cette mesure a été évaluée par les travaux de [Mil, 1993] à travers une expérience qui utilise des sujets humains pour l'évaluation de la similarité entre 30 paires de nom et elle donne une amélioration significative [Thabet et al, 2007].

3.3.2.3 La mesure de Hirst et St-Onge:

Selon [Andon, 2012] le concept de chaines lexicales développé par [Morris et Hirst, 1991] a été adapté par [Hirst et St-Onge, 1998] comme mesure de similarité sémantique qui utilise la structure de WordNet. En effet, cette mesure est basée sur l'idée de [Halliday et Hasan, 1976] que les concepts dans un texte donné ont une forte probabilité de se référer à des concepts déjà mentionnés ou à d'autres concepts reliés, et que des chaines cohésives sont formées par l'enchainement de ces concepts. Voici un exemple de chaine lexicale:

Rome->ville->habitant et manger->plat->légume->aubergine.[Navigli, 2009]

D'après [Zargayouna et Salotti, 2004] toutes les relations dans WodrNet sont prises en considération par la mesure de [Hrist et St-Onge, 1998] ainsi que les liens sont classé comme horizontal (antonyme), bas (sous-classe), haut (partie-de).

Cette mesure est définie selon la formule ci-après:

$$Sim_{Hso} = C - N1 - N2 - k * virage(s1,s2)$$

Où: C et K sont deux constants et la fonction virage(s1,s2) retourne le nombre de changement de direction entre les sens s1 et s2.

3.3.3 Les approches Hybrides:

D'après [Thabet et al, 2007], les approches se basent sur les deux approches citées précédemment qui sont les approches basées sur les arcs et les approches basées sur le contenu informationnel, ces dernières ils sont considérées comme facteur de décision.

3.3.3.1 La mesure de Jiang et Conrath:

A fin de trouver une solution pour le problème de la mesure de Resnik, [Jiac, 1997] a proposé une nouvelle formule qui combine entre les approches basées sur les nœuds et les approches basées sur les arcs par la combinaison entre le contenu informationnel du PPG et à ceux des concepts qu'on veut calculer la similarité.

Selon [Zargayouna et Salotti, 2004] cette mesure prend en compte le nombre d'arcs et elle définie une distance selon cette formule:

$$distance(C1,C2) = CI(C1) + CI(C2) - (2 * CI(ppg(C1,C2)))$$

Donc la nouvelle mesure de [Jiang et Conrath, 1997] est la suivante:

$$Sim(C1,C2) = \frac{1}{distance(C1,C2)}$$

3.3.3.2 La mesure de Leacock et Chodorow:

Selon [Thabet et al, 2007] cette méthode combine entre les approches qui se basent sur les nœuds (CI) et sur les approches qui se basent sur les distances taxonomiques (comptage des arcs), en plus la longueur du plus court chemin entre deux synsets de WordNet est la base de la mesure de [Lec, 1998]. La formule est définie comme suivant:

$$Sim(X,Y) = -\log(\frac{cd(X,Y)}{2 * M})$$

avec:

M: la longueur la plus longue qui sépare le concept le plus en bas du concept racine de l'ontologie.

cd(X,Y) représente la distance en nombre d'arcs du chemin le plus court qui sépare X de Y.

3.4 Conclusion

Dans ce chapitre on a essayé de tracer dans l'état de l'art les différentes mesures sémantiques qui permettent de calculer la distance entre concept et concept, or notre travail consiste à calculer la distance sémantique entre document et document.

Le prochain chapitre sera consacré à l'utilisation des mesures de similarité sémantique entre concepts pour calculer le rapprochement entre les documents.

Chapitre 3 : notre travail

4.1 Introduction :

L'objectif de notre travail est la représentation conceptuelle des documents, en effet l'avantage est de regrouper plusieurs mots dans un seul sens. De ce fait on peut considérer un document contenant le mot « véhicule» comme étant proche à un document contenant le mot « car ». Alors que dans la représentation « sac de mot » cette caractéristique n'existe pas et les deux documents précédents sont considérés disjoints.

Néanmoins, même si un sens n'existe pas dans un autre document, il n'est pas juste d'affirmer qu'il est disjoint, en effet il est judicieux de considérer un document contenant le mot véhicule comme étant proche à un autre document contenant le mot car.

Notre objectif est d'améliorer la représentation conceptuelle en prenant en considération les approches entre les sens, ce qui veut dire ne pas écarter un sens s'il n'existe pas (possible qu'il existe un sens plus proche dans l'autre document).

4.2 Architecture de notre approche

Comme montré dans la figure N°11 l'architecture de notre approche est composée de l'apprentissage et de la classification et les différentes étapes pour classer un document sont énoncées comme suit :

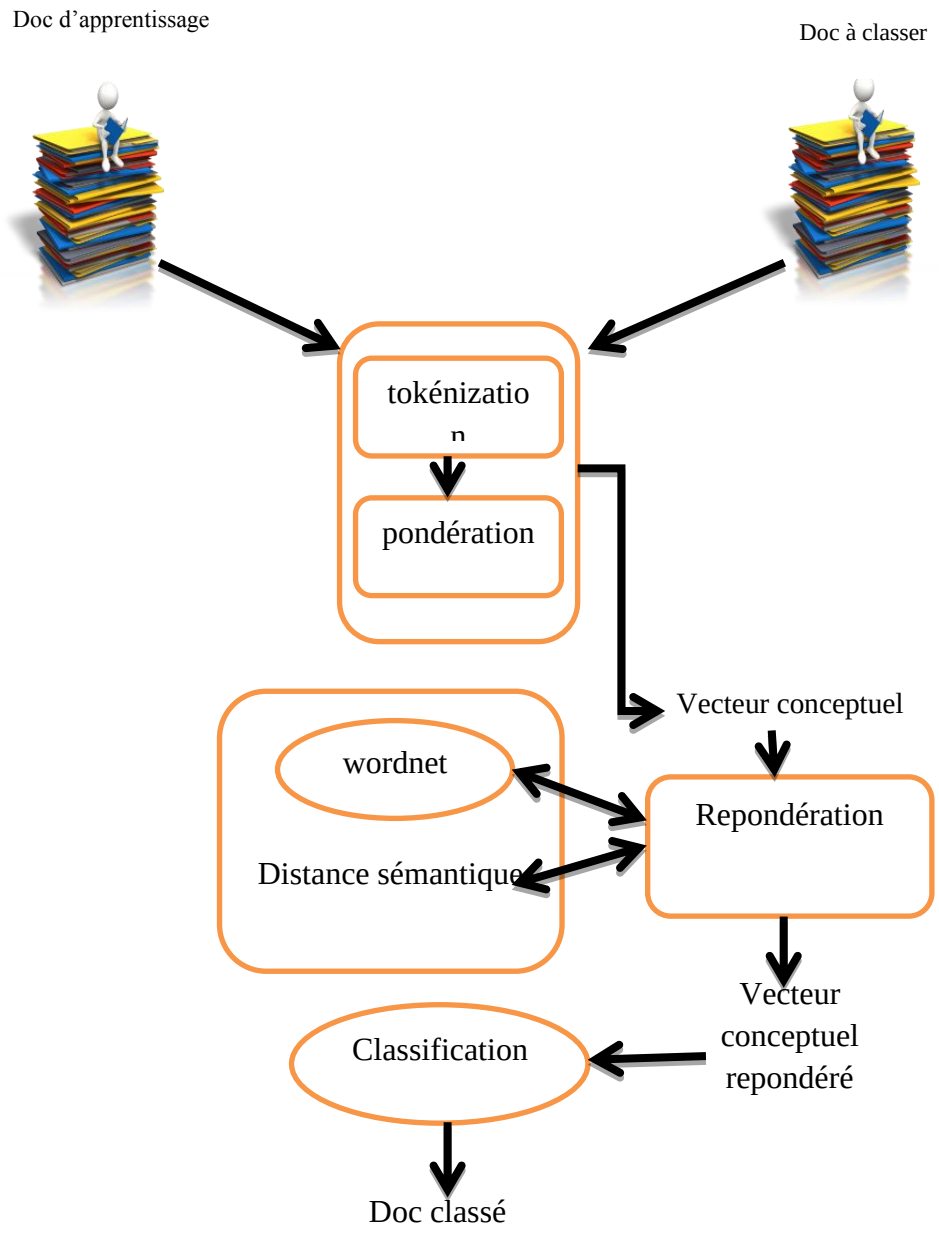


Figure 11-Architecture

4.2.1 Tokenization :

Dans cette approche il s'agit de découper tous les documents en plusieurs unités, la question qui se pose : quoi choisir comme unité ?

Dans notre approche on a choisi de découper le texte en plusieurs phrases. En effet les phrases portent beaucoup plus le sens que le mot tout seul.

Exemple : si on considère le mot : «souris » lui seul on peut comprendre la souris comme étant l'animal, mais si par contre on utilise ce mot dans une phrase le sens du mot sera plus précis : « la souris appartient au périphériques d'entrés » on comprend toute suite qu'on parle d'un composant électronique.

La figure suivante montre le découpage doc/phrase sur les documents de notre collection.

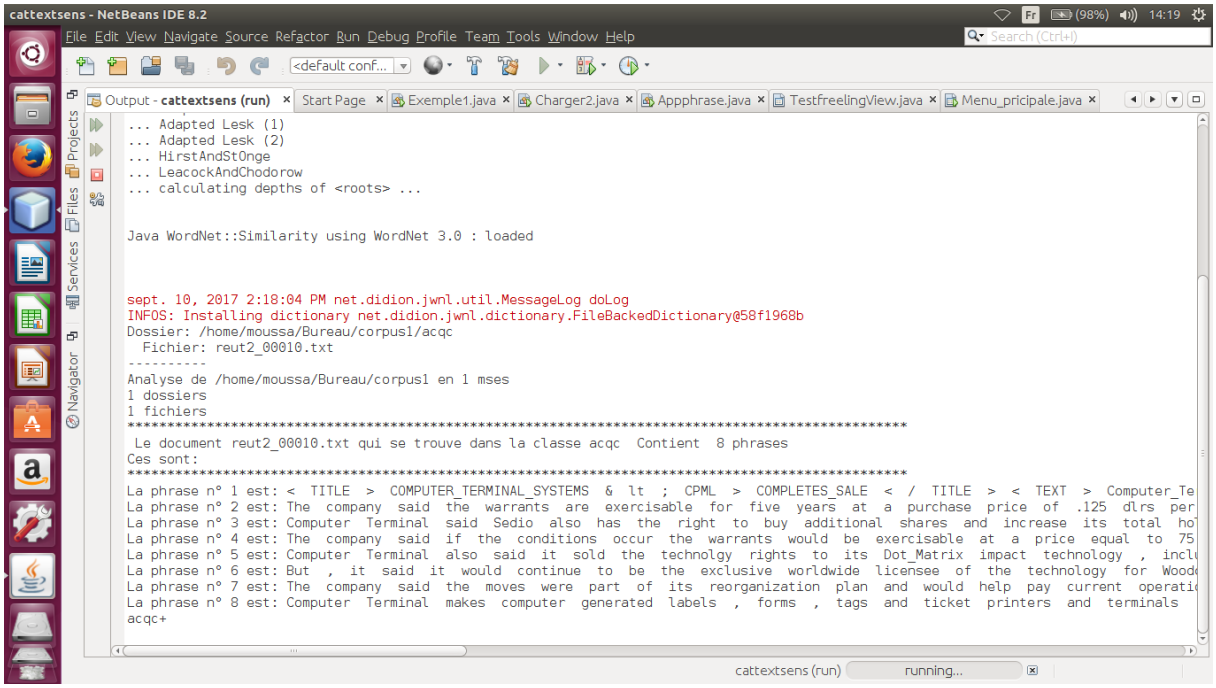


Figure 12-découpage document/phrase

Une fois les phrases repérées une deuxième tokénization s'impose à l'intérieur de chaque phrase pour les découper en plusieurs sens.

Cette étape nécessite d'effectuer un étiquetage grammatical à savoir détecter la caractéristique grammaticale d'un mot donné, c'est à dire de trouver les caractéristiques : verbe, adjectif, nom, cette fonction est réalisée en utilisant la bibliothèque FREELING. Par exemple le mot « plan » dans la figure 13 est un nom.

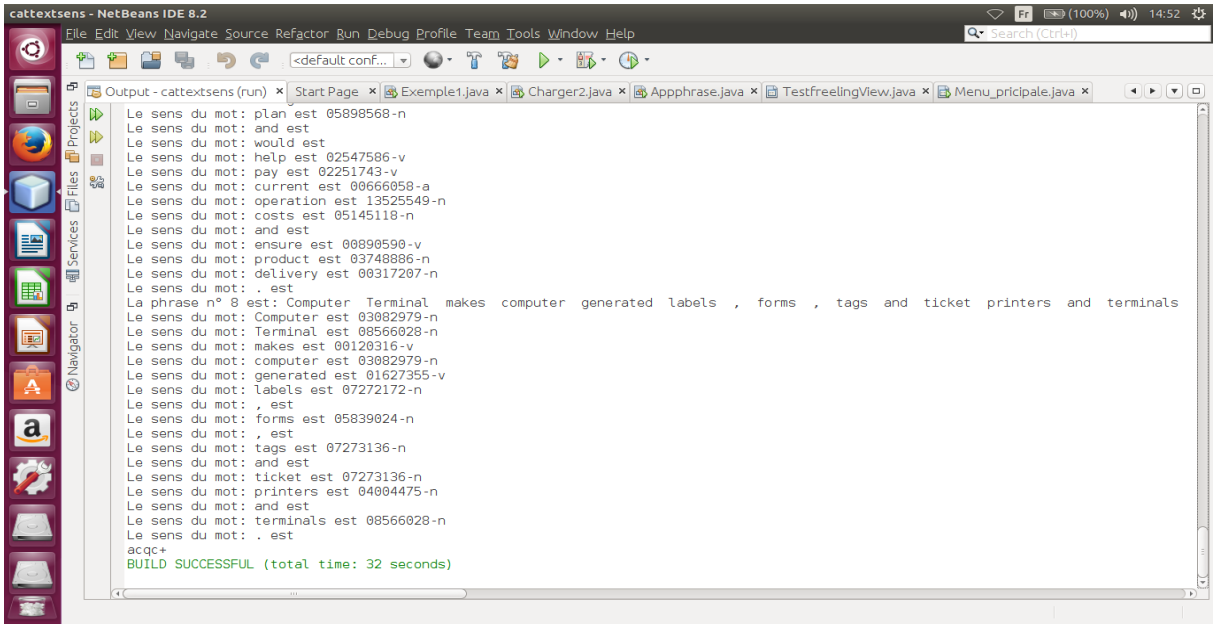


Figure 13-découpage phrase/mot+sens

En effet il est nécessaire de trouver la grammaire avant le sens.

Exemple : le mot « porte » peut être un nom qui signifie l'ouverture de la maison comme il peut être un verbe qui signifie tenir.

D'après la figure 13 l'offset d'un mot se compose de deux parties: le numéro de son sens et la catégorie grammaticale.

4.2.2 **Pondération :**

Dans cette étape il s'agit d'affecter un poids à chaque sens, dans notre travail on a utilisé comme pondération la TF (term frequency) qui consiste au nombre d'occurrence du sens dans un document, ainsi plus le terme est fréquent plus il est important.

Une fois la tokénization et la pondération sont effectuées pour tous les documents d'apprentissage le résultat est la matrice de cooccurrence (doc x sens) comme illustré dans la figure suivante :

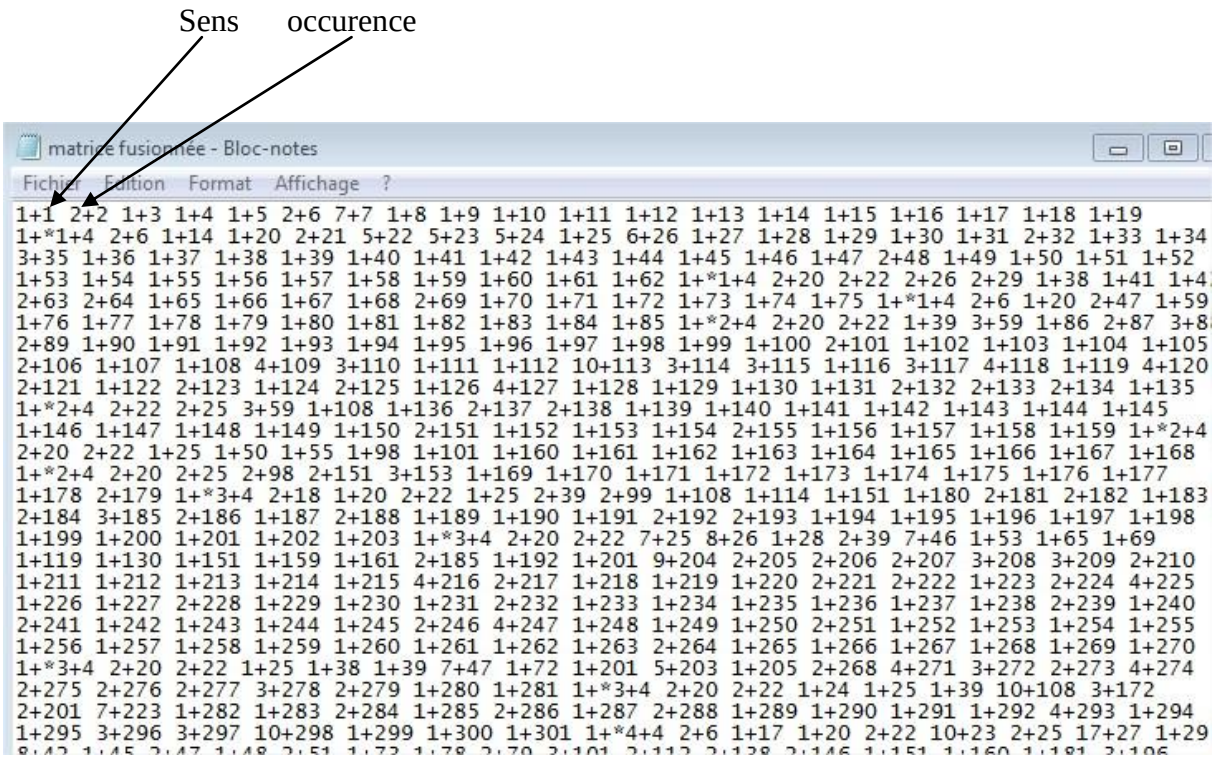


Figure14-matrice cooccurrence doc x sens

4.2.3 Apprentissage et classification :

Dans cette étape il s'agit de trouver la classe du nouveau document, pour cela il est nécessaire de lui appliquer la tokénization/pondération(vu précédemment) ce qui nous donne comme résultat un vecteur conceptuel du document à classer comme illustré dans la figure suivante :

Le document à classer est de la classe earnc et le résultat obtenu est 21 sens, par exemple le sens S3 se répète deux fois dans le document.

S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11	S12	S13	S14	S15	S16	S17	S18	S19	S20	S21
2	1	2	2	1	1	2	1	1	2	2	1	1	1	1	1	1	1	1	1	1

Tableau N°3-vecteur conceptuel du document à classer

Avec les sens S :

s1	06339416-n
s2	06387980-n
s3	01009240-v
s4	08322981-n
s5	10014939-n
s6	00806502-v
s7	13333833-n
s8	00439043-n
s9	00485711-a
s10	13285176-n
s11	10657969-n
s12	06647206-n
s13	08058098-n
s14	00047534-r
s15	02461314-v
s16	00875141-v
s17	00678024-a
s18	08307589-n
s19	07356676-n
s20	00803325-v
s21	13354420-n

Tableau N°4-identification des sens

Les catégories grammaticales pour la langue française sont définies comme suit :

- A:adjective
- C:conjonction
- D:determinant
- N:nom
- P:pronom
- R:adverbe
- V:verbe
- Z:nombre
- W:date

4.2.4 Repondération du vecteur conceptuel du document à classer:

C'est dans cette étape que réside l'essentiel de notre intervention. En effet plusieurs sens du document à classer peuvent ne pas figurer dans les sens des documents d'apprentissage et par conséquence ils ne participeront pas dans la décision.

Prenant un exemple :

DOC 13

La phrase n° 1 est: < TITLE > AM_INTERNATIONAL_INC & lt ; AM > 2ND QTR_JAN 31 < / TITLE > < TEXT > Oper shr loss two cts vs profit seven cts Oper shr **profit** 442,000 vs profit 2,986,000 Revs 291.8 mln vs 151.1 mln Avg shrs 51.7 mln vs 43.4 mln Six mths Oper shr profit nil vs profit 12 cts Oper net profit 3,376,000 vs profit 5,086,000 Revs 569.3 mln vs 298.5 mln Avg shrs 51.6 mln vs 41.1 mln NOTE : Per shr calculated after payment of preferred dividends.

Le mot profit: qui signifie le gain.

DOC 50

La phrase n° 2 est: **Mortgage** Fund Two > said it is making a special distribution of 71.6 cts per exchangeable unit , which includes 67.62 cts from return on capital and 3.98 cts from income gains . Fin phrase 2

Le mot Mortgage: qui signifie le fonds hypothécaire.

La similarité entre les deux sens correspondant à ces deux mots est: 0,5714

Donc il est important de prendre en considération ce rapprochement entre les sens.

Nous proposons de faire une repondération des sens ne figurant pas dans les sens d'apprentissage, en effet il s'agit de comparer le sens inexistant avec tous les sens d'apprentissage et lui affecter une nouvelle pondération comme montre la formule suivante :

Poids(S_i, d_j)=TF(S_i, d_j) x distance sémantique(S_i, S)

Avec S : le sens le plus proche au sens S_i parmi les sens d'apprentissage.

Pour calculer la distance sémantique il existe plusieurs méthodes (voir chapitre 2) dans notre travail on a choisi une des méthodes (wu and palmer, Resnik, Lin, Hirst et St-Onge, Jiang et Conrath, Leacock et Chodorow).

4.2.5 Classement du document :

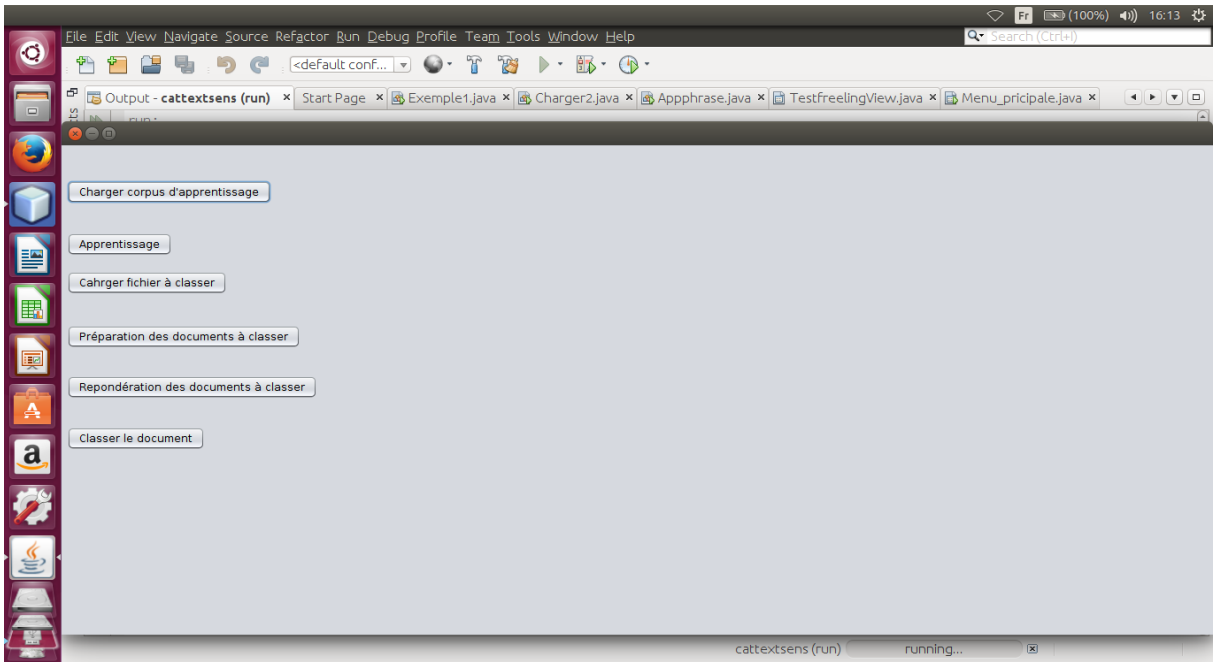
Une fois la repondération effectuée, il s'agit de trouver la classe du document, plusieurs méthodes de classification existent (voir chapitre 1) dans notre travail on a choisi la classification des K plus proches voisins.

En effet il s'agit de calculer la distance du document à classer avec tous les documents classés, finalement le document sera affecté à la classe la plus répondu parmi les K plus proches documents.

La distance entre les documents est calculée en utilisant le produit scalaire d'où la formule est la suivante :

PS (d_1, d_2)=$\sum_{i=1}^{n}$ $poids(sens_i, d_1)$x $poids\ (sens_i, d_2)$

4.3 Figure du programme



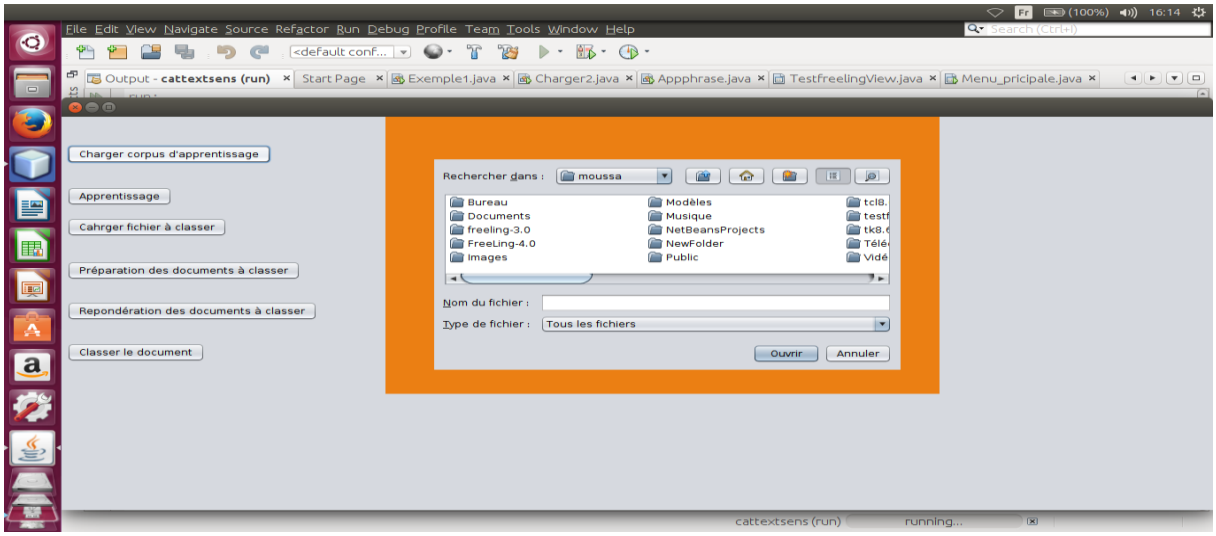


Figure 15-interface de l'application

4.4 Les ressources utilisées :

4.4.1 Description des corpus

Nous avons pris deux corpus le premier est le corpus d'apprentissage composé de huit classes où chaque classe contient quatre documents. Le second est le corpus de test qui comporte un seul document.

4.4.2 Environnement de travail :

Dans notre application nous avons utilisé l'environnement NetBeans open source par Sun sous licence CDDL (Common Development and Distribution License), il permet la prise en charge native de divers langages comme le Java, C, le C++, le JavaScript, le XML, le Groovy, le PHP et le HTML, Python. Il offre toutes les facilités d'un IDE moderne (éditeur en couleurs, projets multi-langage, refactoring, éditeur graphique d'interfaces et de pages Web).

Compilé en Java, NetBeans est disponible sous Windows, Linux, Mac, et possède même une version indépendante des systèmes d'exploitation (requérant une machine virtuelle Java). Un environnement Java Development Kit JDK est requis pour les développements en Java.

NetBeans constitue par ailleurs une plate forme qui permet le développement d'applications spécifiques (bibliothèque Swing (Java)). L'IDE NetBeans s'appuie sur cette plate forme.

4.4.3 La bibliothèque FREELING:

Freeling est une bibliothèque sous C++ compatible avec les langages de programmation les plus réputé tel que : Java, Python, Perl ; qui permet d'analyser des textes sous différentes langues (espagnol, français, anglais, russe, portugais…) en utilisant des techniques comme : détection de la langue, désambiguïsation, étiquetage, parceur, détection des phrases…

Cette bibliothèque inclue Wordnet des langues les plus réputées qui utilise les dictionnaires des sens ainsi les relations qui les connectent te que hyperonymie.

4.4.4 Java:

Nous avons choisis le langage de programmation java qui est orienté objet aux cœurs des méthodes de programmations modernes. Il contient une bibliothèque très riche ainsi une grande quantité de codes réutilisables et un environnement d'exécution qui propose des services tels que la sécurité, la portabilité sur les systèmes d'exploitation et le ramasse-miettes automatique.

Les principaux avantages de java sont : Simplicité, portabilité(l'indépendance de la mise en œuvre), interprétation(lebyte code peut être interprété sur n'importe quelle machine), distribution(accédé depuis le net), fiabilité, multithread(possibilité de programmer en multithread), sécurité(utilisé dans des serveurs), dynamisme, architecture neutre(génération de bytecode exécuté sur différents processeurs).

4.4.5 WordNet :

Selon [Thanh Ngoc Dao&Troy Simpson, 2005]Wordnet est une base de donnée lexicale disponible en line qui traite un très grand nombre de sujets en anglais.

Il existe un wordnet multilingue intitulé :european wordnet possédant les mêmes structures que celle de wordnet.

Wordnet établie les liens entre les parties de langage (nom, verbe, adverbe, adjectif). La plus petite unité dans wordnet est appelée « synset » qui représente un sens spécifique pour un mot, elle inclus le mot, ses synonymes ainsi que ses explications. Chaque sens d'un mot constitue un Synset à part.

Chaque Synset possède un « Gloss »définissant le concept qui le représente.

Exemple :nigth, nigthing et dark constitue un Synset ;son gloss : « le temps avant le levé de soleil et après le couché de soleil quand il fait sombre dehors ».

Les Synsets sont reliés les uns aux autres avec des relations sémantiques (hypernym, hyponym pour les noms et hypernym, troponym pour les verbes)qui constituent :genre-de(holonymy) et partie-de(meronymy).

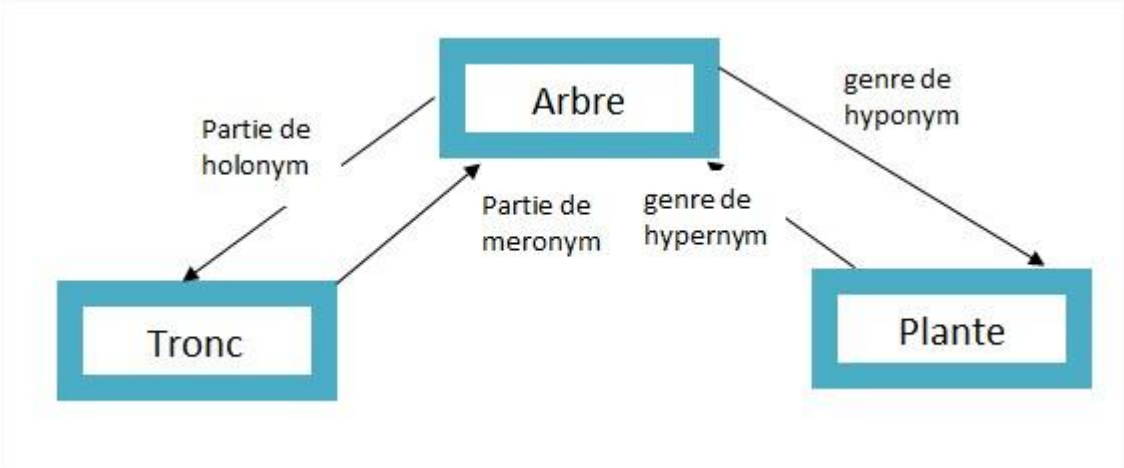


Figure16-relations sémantiques de l'arbre

Malcolm Crone et Troy simpson ont développé un Framework open source nommé Wordnet.net sous C#pour Wordnet1.6 ensuite 2.0 et 2.1.

4.4.6 La bibliothèque JWS(java wordnet similarity)

Cette bibliothèque de similarité permet de fournir un outil pour calculer la similitude à la fois entre les mots et les phrases en utilisant les fonctionnalités de Wordnet, dans le cas des mots elle permet de trouver les relations entre eux, tandis que dans les cas des phrases elle utilise des algorithmes qui calculent les similitudes entre elles.

4.5 Conclusion

Pour conclure ce chapitre on peut dire qu'on a expliqué l'architecture de notre approche qui se compose de la tokénization, la pondération, l'apprentissage et la classification, la repondération et en final le classement. Notre expérimentation produit plusieurs niveaux de découpage en partant du document puis la phrase puis le mot et son sens, et nous avons montré que l'étiquetage grammatical représente une opération très importante avant de chercher le sens.

La repondération de la matrice fusionnée (figure N°14) permet d'éviter d'ignorer les sens inexistants mais qui se rapprochent du sens du document à classer, le document N°13 à classer contient le mot « profit » dans la phrase N°1, ce sens n'existe pas dans les document d'apprentissage, donc après repondérations on a trouvé le document N°50 qui contient le mot « Mortgage » dans la phrase N°2 avec une similarité de 0.5714, soulignant que les mot profit et hypothèque se rapprochent.

Enfin on a exposé les ressources utilisées dans notre travail qui sont : Netbeans, la bibliothèque Freeling, java, wordnet, java wordnet similarity.

Conclusion générale

A la fin de ce mémoire il est nécessaire de rappeler d'une façon globale ce qui était abordé, au début du premier chapitre nous avons évoqué les définitions de la catégorisation pour énoncer ses différents types et par la suite les étapes à suivre pour effectuer la C.T (indexation, chaix du classificateur, évaluation). La première étape permet de réduire le nombre des termes pour garder les plus importants afin de gagner en espace de stockage et gagner en temps d'exécution. L'étape suivante sera la plus importante car elle permet de choisir le type du classificateur et enfin l'évaluation du model qui permet de déterminer sa qualité.

Le deuxième chapitre nous a permis au début de définir la similarité pour énoncer les principaux approches qui existent (arcs (Wu et Palmer, Zargayouna et Salotti, Thabet et al, Rada et al), nœuds (resnik, lin, Hirst et St-Onge), hybride(Jiang et Conrath, Leacock et Chodorow)).

Dans le troisième chapitre qui est la partie essentiel de notre travail, au début on a présenté l'architecture de l'approche qui comporte la tokénization, la pondération, l'apprentissage et la classification, la repondération et enfin le classement de notre document. Ensuite on a évoqué les ressources utilisées qui sont Netbeans, la bibliothèque Freeling, java, wordnet, java wordnet similarity.

Néanmoins notre travail sera complété dans le futur en abordant les points suivants :

- Tester notre approche pour différents corpus.
- Essayer d'autres mesures de pondération à part la TF.
- Comparer les différentes mesures de similarité à part WuandPalmer

Bibliographie

[Andon, 2012]	A.Tchechmedjiev. Mesures de Similarité Sémantique Locales et Algorithmes Globaux pour la Désambiguisation. Conférence Paper. June 2012.
[Elsa Negre 2013]	Elsa Negre. Comparaison de textes 2013
[Groin&Al, 2009]	
[Halliday et Hasan, 1976]	HALLIDAY, M. A. et HASAN, R. (1976). Cohesion in English. Longman Group Ltd, London, U.K.
[Hrist et St Onge, 1998]	HIRST, G. et ST ONGE, D. D. (1998). Lexical chains as representations of context for the detection and correction of malapropisms. WordNet : An electronic Lexical Database. C. Fellbaum., pages 305 à 332. Ed. MIT Press.
[Jalam, 2003]	Apprentissage automatique et catégorisation de textes multilingues
[Jiac, 1997]	J. Jiang et D. Conrath. Semantic similarity based on corpus statistics and lexical taxonomy. In Proceedings of International Conference on Research in Computational Linguistics, Taiwan, 1997.
[LAHLOU OUCHIHA, 2016]	classification supervisée de documents université du québec en outaouais
[Lec, 1998]	C. Leacock et M. Chodorow. Combining Local Context and WordNet Similarity for Word Sense Identification. In WordNet: An Electronic Lexical Database, C. Fellbaum, MIT Press, 1998.
[Lee, 1993]	J.H.Lee, M.H.Kim et Y.J.Lee. Information Retrieval Based on Conceptual Distance in IS A Hierarchy. Journal of Documentation 49, pp. 188 à 207, 1993.
[Lin, 1998]	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.
[Mathieu Stricker,2004]	Réseaux de neurones pour le traitement automatique du langage
[Morris et Hirst, 1991]	MORRIS, J. et HIRST, G. (1991). Lexical cohesion computed by thesaural relations as an indicator of the structure of text. Comput. Linguist., 17(1):21 à 48.
[Navigli, 2009]	NAVIGLI, R. (2009). Word sense disambiguation: A survey. ACM Comput. Surv., 41(2):10 :1 à 10 :69.
[Ph. PREUX, 2011]	Fouille de données Notes de cours Philipe PREUX Université de Lille 3
[Rada 89]	R. Rada, H. Mili, E. Bichnell et M. Blettner, Development and application of a metric on semantic nets. IEEE Transaction on Systems, Man, and Cybernetics: pp 17à 30. 1989.
[Res, 1995]	P. Resnik. Using information content to evaluatesemantic similarity in taxonomy. In Proceedings of 14th International Joint Conference on Artificial Intelligence, Montreal, 1995.
[Res, 1999]	P. Resnik. Semantic similarity in a taxonomy: An information based measure and its application to problems of ambiguity in natural language. Journal of Artificial Intelligence Research, 11:95à130, 1999.
[Sebastian, 2010]	Approches textuelles pour la catégorisation et la recherche de documents manuscrits en ligne
[Sebastiani, 2002]	Machine Learning in Automated Text Categorization

[S.L. Ting, W.H. Ip, Albert H.C.Tsang ,2011] Is Naïve Bayes a Good Classifier for Document Classification?-
Department of Industrial and Systems Engineering, The Hong Kong Polytechnic

[Thabet et al, 2007] S.Thabet, B.Y.Boutheina et K.Mellouli. Une extension de mesure de similarité entre les concepts d'une ontologie.Article scientifique.2007.

[Thanh Ngoc Dao&Troy Simpson, 2005] Measuring Similarity between sentences

[Tsoumakas and Katakis, 2007] Multi-Label Classification . Aristotle University of Thessaloniki, Greece

[Tversky, 1977] A. Tversky. Features of similarity. Psychological Review, 84(4) :327 à 352, 1977. 106

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

[Zargayonna,2005] H.Zargayouna. Indexation sémantique de document XML, thèse pour l'attentiez du grade de docteur, chair pari XI UFR scientifique d'Orsay. page 160.

[Zargayouna et Salotti, 2004]Z.Haifa et S.Sylvie. Mesure de similarité sémantique pour l'indexation de documents semi structurés. Article January 2004.

Résumé

Ce mémoire s'inscrit dans le cadre des problèmes liés à la repondération des textes, enfaite il existe plusieurs méthodes dont chacune d'elles possède des avantages et des inconvénients et on a montré qu'il est judicieux d'utiliser la méthode conceptuelle qui permet de considérer les sens des mots à l'intérieur des phrases parce qu'un mot est plus précis dans une phrase que si il était isolé. Ainsi la tokenization qui produit des phrases est plus efficace que celle qui produit des sacs de mots ou des n-grams.

Mots clé

Catégorisation des textes-contenu informationnel-plus petit généralisant-similarité-synset-sens-distance.

Abstract

This memory falls within the scope of the problems related to the reweighting of texts. There are several methods, each of which has advantages and disadvantages, and it has been shown that it is judicious to use the conceptual method which allows to consider the meaning of words within sentences because a word is more precise in a sentence than if it were isolated. So the tokenization that produces sentences is more effective than the one that produces bags of words or n-grams.

Key words

Text categorization -information content- smaller generalizing -similarity- synset- sense-distance

ملخص

تندرج هذه الأطروحة في إطار المشاكل المتعلقة بإعادة ترجيح النصوص، وهناك عدة طرق، لكل منها مزايا وعيوب، وقد تبين أنه من الحكمة استخدام الأسلوب المفاهيمي الذي يسمح للنظر في معنى الكلمات داخل الجمل لأن الكلمة أكثر دقة في جملة منها إذا كانت معزولة.

وبالتالي فإن الرمزية التي تنتج الجمل هي أكثر فعالية من تلك التي تنتج أكياس من الكلمات أو n- غرام.

الكلمات المفتاحية

تصنيف النصوص-محتوى المعلومات-أصغر تعميم-التشابه-synset-المعنى- المسافة.