4th International Conference on Science, Technology and Management
India International Centre, New Delhi (ICSTM-16)
15th May 2016, www.conferenceworld.in ISBN: 978-81-932074-8-2

# L'INDEXATION SEMANTIQUE A L'AIDE D'UNE ONTOLOGIE D'ENTREPRISE

**Anass Mamouny[1], Mostafa Hanoune[2]**

[1, 2] *Doctoral Studies: Mathematics, Computer Science, and Information Technologies Department*
*Hassan II Mohammedia-Casablanca University, Ben M'sik School of Sciences*
*BP 7955 Casablanca, Morocco*

**ABSTRACT**

*Dans cet article, nous proposons un modèle semi-automatique pour indexer les documents à l'aide d'une ontologie nommée EntrOnto, cette ontologie décrit l'environnement intérieur et extérieur de l'entreprise dans le contexte de la démarche de l'intelligence économique. Nous commençons par un état de l'art sur les méthodologies de l'indexation sémantique. Ensuite, nous présentons le scenario de notre modèle proposé. Enfin, nous expliquons à l'aide d'un exemple la procédure pratique de l'indexation.*

***Keywords: Indexation sémantique, Ontologie d'entreprise, TAL, Système de recherche d'information.***

## I. INTRODUCTION

L'indexation est une étape primordiale dans la recherche d'information permettant d'extraire d'un document ou d'une requête une liste de termes (ou groupe de termes) appelés descripteurs ou indexes. Généralement, les descripteurs sont assortis de poids représentant le degré de leur représentativité sémantique de document. Cette opération peut être réalisée d'une manière : manuelle, semi-automatique, ou automatique. Dans l'indexation manuelle, les documents sont traités « manuellement » de la part d'une personne expérimentée dans le domaine de recherche [1]. Grâce à l'intervention de spécialistes, l'indexation manuelle construit une relation performante entre les documents et leurs indices. Cela permet de définir un SRI effectif qui retourne des choix précis en réponse au lancement d'une requête. Le désavantage majeur de cette méthode est le grand effort qu'il faut fournir pour définir les représentations des documents. En outre, quoique efficace, cette méthode d'indexation est intrinsèquement affecté par les choix subjectifs de l'intervenant humain ce qui peut créer des indexations très différentes pour un même document ou pour des documents semblables. Bien que, l'indexation semi-automatique est un processus automatisé est adopté pour définir les choix potentiels répondant à la requête lancée par l'utilisateur [2]. Un seul choix est choisi de l'ensemble des choix par un spécialiste. L'indexation semi-automatique s'appuie sur la création d'un thesaurus auxiliaire contenant les relations liant les champs sémantiques à l'aide de rapports logiques prédéfinis. Les techniques d'indexation automatique sont utilisées pour construire le dictionnaire des champs sémantiques avant l'intervention d'un spécialiste pour effectuer un choix pertinent parmi les termes retenus. Par contre, dans l'indexation automatique, le choix est effectué à l'aide d'un processus automatisé sans intervention humaine [1]. L'indexation automatique permet une réduction considérable de l'effort requiert pour indexer un

document. Cependant, l'efficacité de l'indexation peut être affectée par les algorithmes adoptés pour automatiser le choix des représentations. Cela justifie les efforts de recherche fournis pour améliorer ce mode d'indexation. De plusieurs étapes existent dans un processus d'indexation automatique. Par exemple, l'élimination de termes insignifiants et la pondération des éléments sont deux phase qui précèdent la construction des représentations et qui affectent la vitesse et la performance de la RI. L'indexation automatique sera étudiée plus tard dans ce travail.

## II. ETAT DE L'ART

L'indexation sémantique consiste à associer à chaque terme un champ notionnel prédéfini par l'utilisateur. Par exemple, le terme « vert » peut référer à la couleur, ou à un mouvement politique et sociétal. Cela nécessite la construction d'un catalogue ou une ontologie contenant les concepts décrivant tous les termes potentiels pour les associer à leurs concepts. On parle alors du processus de désambiguïsation des termes.

Voorhees [3], Katz [4], et Uzuner [5], ont proposé une approche d'indexation basée sur le sens des mots d'un document en exploitant la base WordNet [6] pour récupérer les synsets correspondants. Si un mot correspond à plusieurs synsets, il est qualifié d' « ambigu ». Dans ce cas, on explore son contexte local sur le document pour déterminer le sens approprié. On compare les mots communs entre le voisinage de chaque synset dans WordNet et son contexte local. Toutefois, les expérimentations montrent que cette approche obtient des performances inférieures comparativement à une approche classique utilisant simplement les mots clés. Gonzalo [7] a réussi a amélioré les performances de RI avec un pourcentage de 29 %. Cependant, son approche est basée sur la désambiguïsation manuelle des documents.

Baziz et ses collègues [8] ont développé une méthode d'indexation sémantique des documents. L'approche de désambiguïsation proposée s'appuie sur les mesures de similarité sémantique. L'hypothèse fondamentale de cette approche dit que le sens le plus adéquat d'un terme ambigu est celui qui est le plus sémantiquement proche des autres termes contenus dans son contexte. Les synsets correspondants d'un terme ambigu jouent le rôle des concepts candidats. On calcule le degré (score) de similarité de chaque concept candidat par rapport aux autres concepts du contexte, et le candidat possédant le score maximal est retenu. Une fois les concepts désambiguïsés, ils sont utilisés pour représenter les documents et les requêtes dans un réseau conceptuel. Dans ce réseau, les liens entre concepts sont pondérés par les valeurs de leur similarité sémantique.

Les mesures de similarité utilisées dans cette approche sont la mesure de Resnik [9], La Mesure de Leacock et Chodorow [10], et la mesure La mesure de Lin [11].

La mesure de Rensik est basée sur le contenu informationnel du concept le plus spécifique qui les subsume (« least common subsumer »). Cette mesure considère que deux concepts sont sémantiquement proches si la quantité d'information qu'ils partagent (le contenu informationnel de leur parent commun) est importante. Elle est définie par :

$$\text{Sim Rensik}(C1,C2) = IC(C)$$

Avec C est le concept le plus spécifique qui subsume C1 et C2, et IC est le contenu informationnel du concept C. Le contenu d'information (IC) d'un concept est estimé en calculant sa fréquence dans un large corpus. Il est défini par :

$$IC(C) = -\log(P(concept))$$

La fréquence d'un concept C dans la hiérarchie, inclut la fréquence de tous ces descendants puisque une occurrence ajoutée à un concept est aussi ajoutée aux concepts qui le subsument.

La mesure de Lin est aussi basée sur la notion de contenu informationnel IC d'un concept C, mais il intègre le contenu informationnel de C1 et de C2. Elle est définie par :

$$Sim\ (C1,C2)=(2*IC(C))/(IC(C1)+IC(C2))$$

La mesure de Leacock et Chodorow est basée sur la notion de chemin. Le plus court chemin length(C1,C2) entre deux concepts C1 et C2 est celui qui comprend le plus petit nombre de noeuds intermédiaires. Cette valeur est inversement proportionnelle à la profondeur maximale de l'arbre notée D qui représente la taille du plus long chemin de la feuille au nœud racine (root) dans la hiérarchie. Cette mesure est définie comme suit :

$$Sim\ LCH\ (C1,C2) = \max[-\log(length(C1,C2)/(2.D))]$$

## III. MODÈLE PROPOSÉ

Dans le modèle proposé dans la Figure 1, on adopte une méthode semi-automatique appelé aussi l'indexation supervisée. Nous confirmons que l'intervention d'un utilisateur expert dans le domaine de traitement de documents est indispensable, précisément dans le choix des indexes finaux. Cependant, on recommande d'employer les outils de TAL pour automatiser certaines tâches. Les indexes finaux sont représentés par des objets d'une ontologie nommée EntrOnto que nous avons déjà créé pour modéliser l'environnement intérieur et extérieur de l'entreprise. Ces indexes peuvent être des identifiants de concepts, ou des termes utilisés dans la définition des instances, des attributs, et des relations.

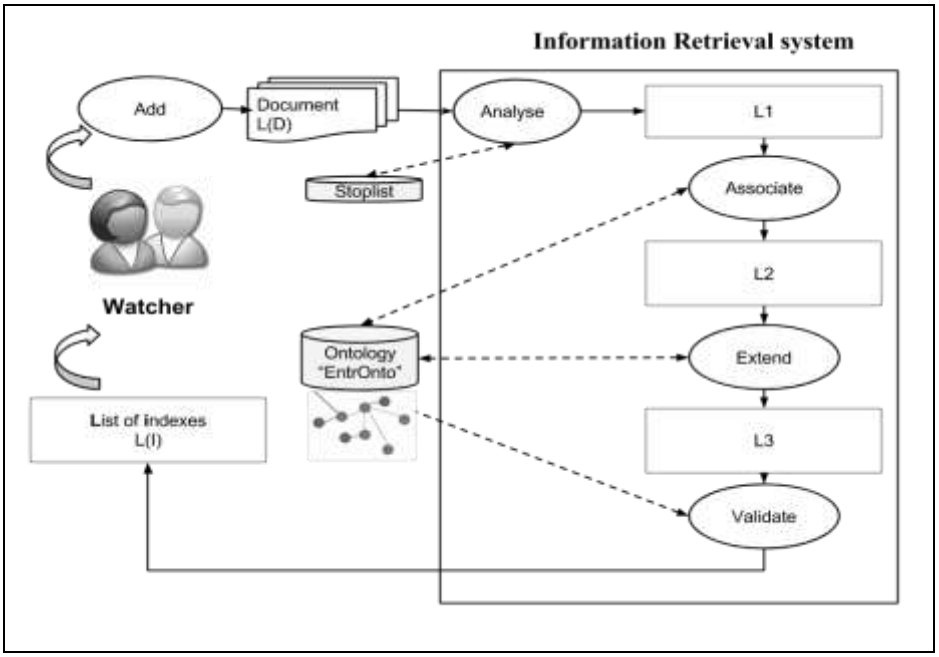


**Fig. 1 Processus d'indexation des documents avec EntrOnto**

Selon notre modèle, le veilleur charge un document D (ou un ensemble des documents E(D)) dans le SRI. L(D) est la liste de tous les termes de document D, tel que :

$$D = L(D) \text{ avec } D \in E(D)$$

Un mot vide (stop word) est un mot non significatif figurant dans un texte. On l'oppose à mot plein ou mot lexical. Ces mots vides sont des mots qui sont tellement communs qu'il est inutile de les indexer ou de les utiliser dans une recherche. En français, Ces mots sont souvent des prépositions ("de", "à"), prénom ("aucun", "tout", "on"), certains adverbes ("ailleurs", "maintenant"), adjectifs ("certain", "possible"), etc. En anglais, des mots vides pourraient être « he », « is », « at », « which », « on », etc. Un anti-dictionnaire (stoplist) est une base de termes qualifiés vides qu'on peut utiliser dans l'analyse des textes et le filtrage des mots qu'on ne veut pas garder dans un document. Le traitement lié à une stoplist est très simple. Quand on rencontre un mot dans un texte, on doit d'abord examiner s'il apparaît dans cette liste. Si oui, on l'élimine.

La fonction « Analyse » dans notre modèle consiste à élaguer les termes vides grâce à un anti-dictionnaire (stoplist). Le système retient une liste des termes candidats d'être dans les indexes de documents (L1), Tel que :

$$L1 = L(D) - L(\text{Termes Vides})$$

La fonction « Associate » dans notre modèle SRI compare chaque terme candidat de L1 avec les termes de l'ontologie EntrOnto. Il ne suffit pas de chercher exactement le terme tel qu'il est parmi les termes d'EntrOnto, mais il faut prendre en considération aussi les équivalents sémantiques (synonymes) et les dérivés linguistiques (nom, adjectif, verbe) dans la recherche. En d'autres termes, le SRI doit rapprocher sémantiquement et linguistiquement entre les termes candidats et les termes de l'ontologie pour déterminer les descripteurs les plus pertinents de document, on parle alors de la désambiguïsation des termes.

Kahan [12,13], Paralic [14], Vallet [15] ont travaillé sur des approches permettant d'identifier manuellement les éléments de l'ontologie correspondants aux termes candidats de document. Cette opération doit être réalisée par un expert de domaine capable d'associer à chaque terme l'élément le plus approprié dans l'ontologie. Cette méthode est couteuse en termes de charge et de temps et elle reste subjective. On préfère d'utiliser une approche permettant d'automatiser l'identification des éléments de l'ontologie qui apparaissent dans le document.

Le modèle proposé permet d'affecter un poids (TF*IDF) à chaque terme existant pour déterminer son importance dans le réseau sémantique de document, et les termes ayant un poids important sont retenus dans une deuxième liste (L2), Tel que:

$$L2 = L1 - L(\text{Termes} \notin \text{EntrOnto})$$

La fonction « Extend » dans notre modèle SRI permet d'étendre la liste (L2) par d'autres termes qui sont fortement liés à la liste retenue à travers les relations d'EntrOnto. Le SRI ajoute à la liste (L2) seulement les termes connectés directement par l'une des relations d'EntrOnto aux termes de (L2). On propose que la longueur de la relation ne doit pas dépasser N=1 afin d'avoir des indexes pertinents. La liste résultat est la troisième liste (L3), tel que :

$$L3 = L2 + L(\text{Termes connectés par EntrOnto})$$

Enfin, la fonction « Validate » permet au veilleur d'analyser (L3) pour valider la liste des indexes finaux de document L(I), tel que :

$$L(I) = L3 - L(\text{termes non validés})$$

Les indexes de L(I) seront utilisés dans la sélection pertinente des sources de données répondants aux requêtes de recherches d'information saisies dans le SRI.

## IV. EXEMPLE PRATIQUE

L'exemple de cette section sert à mettre en pratique la procédure que nous avons proposée pour indexer les documents en se basant sur l'ontologie EntrOnto. L'exemple est un extrait d'un article téléchargé depuis le web. Le choix de l'article est arbitraire, ce qui nous intéresse est de montrer comment pratiquer la procédure étape par étape.

La première étape est l'étape d'analyse de document pour élaguer les mots vides grâce à un anti-dictionnaire. L'Université de Neuchâtel de Suisse offre un ensemble de stoplist dans différentes langues sur leur site : stoplist anglais (571 termes), stoplist francais (463 termes).

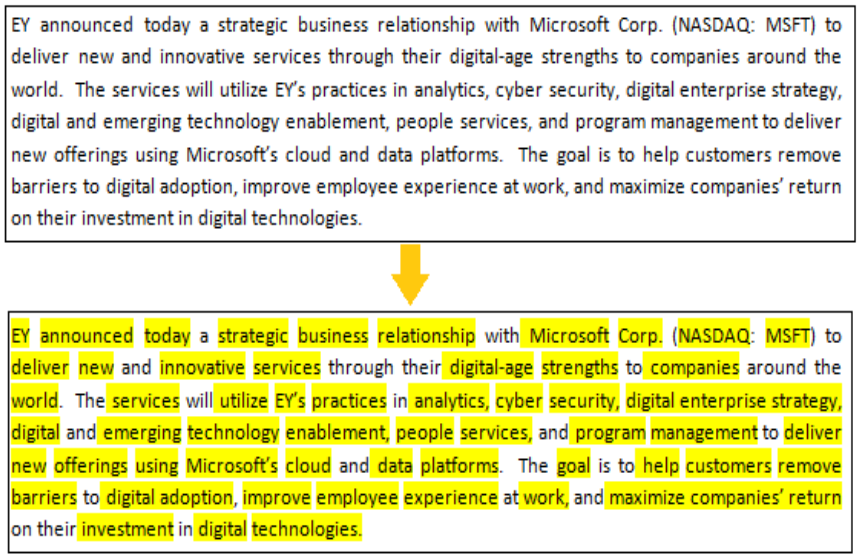


**Fig. 2 Exemple d'extrait article téléchargé depuis le web**

Dans la Figure 2, on a éliminé les termes dite vides, et on a marqué le reste des termes pour construire la liste des termes candidats L1. L1 est composé de 62 termes au lieu de 86 termes de L(D).

Nous avons utilisé l'analyseur de Apache Lucene [16].Cet analyseur contient déjà des filtres pour tout indexer en casse basse, ainsi qu'un StopFilter qui utilise une liste de mots (tokens), les stopwords, qui sont trop courant dans les textes. Par défaut, c'est une liste de mots en anglais définie en interne qui est utilisée. Mais il permet d'y fournir une autre stoplist en paramètre.

Ensuite, nous avons utilisé l'étiqueteur TreeTagger [17], aussi appelé positionneur morphosyntaxique pour obtenir la nature de chacun des termes présents dans le corpus (adjectif, verbe, nom, etc). TreeTagger a été développé par Helmut Schmid dans le ICLUS (Institute for Computational Linguistics of the University of Stuttgart), il repose sur la construction d'un arbre de décision binaire pour estimer les probabilités d'obtenir un certain rôle syntaxique.

Puis, nous avons utilisé l'analyseur syntaxique Syntex [18] pour extraire l'ensemble des syntagmes (unités syntaxiques et sémantiques) de chaque terme de (L1). Syntex permet d'identifier des relations de dépendances entre mots (sujet, complément d'objet, épithète, etc).

Le résultat de l'analyse se présente sous la forme d'un réseau de dépendance, dans lequel chaque syntagme extrait est relié à sa tête et à son expansion syntaxique. Les syntagmes d'un terme peuvent être sous la forme maximale composée de toutes les expansions liées au terme, ou sous la forme réduite qui correspond aux différents syntagmes pour lesquels les expansions sont successivement supprimées. L'intérêt de considérer les différentes formes sous lesquelles se décompose un terme pour les rechercher parmi les éléments de l'ontologie.

Enfin, on a cherché les termes retenus de (L1) parmi les termes d'EntrOnto, et on a trouvé 7 termes présents dans les instances, 4 termes présents dans les concepts, 2 termes présents dans les attributs. 6 termes parmi ces termes sont cités plus d'une fois :

- EY#2
- Microsoft#2
- services#3
- digital#3
- companies#3
- technolgy#2

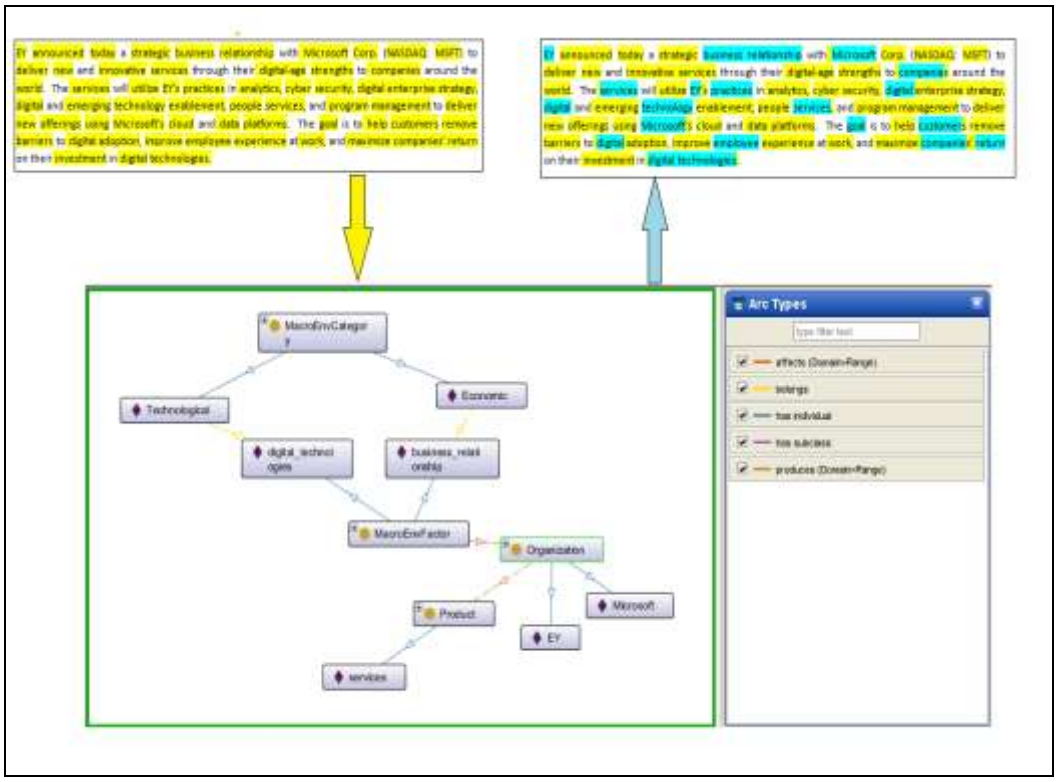


**Fig. 3 Exemple d'annotation de document avec EntrOnto**

Chaque terme de (L2) est relié à d'autres termes à travers l'ontologie EntrOnto. Par exemple, le terme EY instance de la classe Organization est relié avec : le terme Assurance instance de la classe ActivitySector par la relation work_in, le terme Auditing_services instance de la classe Product par la relation produces, et le terme

Ernst_Young_Zurich instance de la classe Unit par la relation consists_of. La Figure 4 montre les relations de terme EY dans EntrOnto :

```
digraph g {
      "Product" -> "Auditing_services" [label="has individual"]
      "EY" -> "Assurance" [label="work_in"]
      "EY" -> "Ernst_Young_Zurich" [label="consists_of"]
      "Organization" -> "Microsoft" [label="has individual"]
      "EY" -> "Auditing_services" [label="produces"]
      "Unit" -> "Ernst_Young_Zurich" [label="has individual"]
      "ActivitySector" -> "Assurance" [label="has individual"]
}
```

**Fig. 4 Exemple d'expansion des indexes avec EntrOnto**

De la même manière, L2 est étendu de 6 termes à 19 termes pour construire une liste L3. Après une vérification manuelle de L3, on a retenu 12 termes pour les indexes de L(I). La liste suivante L(I) montre des exemples des indexes à relier avec le document.

URI : http://www.semanticweb.org/mamouny/ontologies/2015/8/EntrOnto#EY
URI : http://www.semanticweb.org/mamouny/ontologies/2015/8/EntrOnto#Miscrosoft
URI : http://www.semanticweb.org/mamouny/ontologies/2015/8/EntrOnto#digital_technologies
URI : http://www.semanticweb.org/mamouny/ontologies/2015/8/EntrOnto#Auditing_services
URI : http://www.semanticweb.org/mamouny/ontologies/2015/8/EntrOnto#Assurance
URI : http://www.semanticweb.org/mamouny/ontologies/2015/8/EntrOnto#Ernst_Young_Zurich

## Summary

Nous avons pu définir une approche semi-automatique pour indexer les documents en utilisant l'ontologie EntrOnto, ces indexes peuvent être des identifiants de concepts, ou des termes utilisés dans la définition des instances, des attributs, et des relations. Notre approche permet d'accélérer la fonction de comparaison entre les requêtes de recherche et les indexes de documents dans la démarche de l'intelligence économique au sein des entreprises. Concernant nos perspectives, nous envisageons de développer une plateforme logicielle qui traduit notre modèle théorique de SRI à un modèle pratique. Cette plateforme doit avoir une interface conviviale et facile à utiliser par les veilleurs et les décideurs des entreprises.

## REFERENCES

[1] S. Jacques, Indexation manuelle et automatique : une évaluation comparative basée sur un corpus en langue française (2000).

[2] K. Laurent, Indexation semi-automatique de textes : thésaurus et transducteurs. (2009)

[3] E.M. Voorhees, Query expansion using lexical-semantic relations, in: Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '94. Springer-Verlag New York, Inc., New York, NY, USA, pp.61–69.( 1994)

[4] B. Katz, O. Uzuner, D. Yuret, Word Sense Disambiguation For Information Retrieval. (1998)

[5] O. Uzuner, B. Katz, D. Yuret, Word Sense Disambiguation for Information Retrieval, in: Proceedings of the Sixteenth National Conference on Artificial Intelligence and the Eleventh Innovative Applications of Artificial Intelligence Conference Innovative Applications of Artificial Intelligence, AAAI '99/IAAI '99. American Association for Artificial Intelligence, Menlo Park, CA, USA, p. 985–. 1999.

[6] A. Miller George, WordNet: A Lexical Database for English. Communications of the ACM Vol. 38, No. 11: 39-41 (1995).

[7] J. Gonzalo, F. Verdejo, I. Chugur, J. Cigarrin, Indexing with WordNet synsets can improve text retrieval. pp. 38–44. (1998)

[8] M. Baziz, M. Boughanem, N. Aussenac-Gilles, Conceptual Indexing Based on Document Content Representation, in: Crestani, F., Ruthven, I. (Eds.), Context: Nature, Impact, and Role, Lecture Notes in Computer Science. Springer Berlin Heidelberg, pp. 171–186. (2005)

[9] P. Resnik, Semantic Similarity in a Taxonomy: An Information-Based Measure and its Application to Problems of Ambiguity in Natural Language. (1999)

[10] C. Leacock, M. Chodorow, Combining Local Context and WordNet Similarity for Word Sense Identification, in: WordNet: An Electronic Lexical Database. MIT Press. (1998)

[11] D. Lin, An Information-Theoretic Definition of Similarity, in: Proceedings of the Fifteenth International Conference on Machine Learning, ICML '98. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, pp. 296–304. (1998)

[12] L-R. Khan, Ontology-based Information Selection, Phd Thesis, Faculty of the Graduate School, University of Southern California. August (2000)

[13] L-R. Khan, Retrieval effectiveness of an ontology-based model for information selection. TheVLDB Journal 13:71–85. (2004)

[14] J. Paralic, I. Kostial, « Ontology-based Information Retrieval », Information and Intelligent Systems, (p. 22-28). (2003)

[15] D. Vallet, M. Fernández, P. Castells, « An Ontology-based Information Retrieval Model », ESWC (p. 455-470). (2005)

[16] Site web: https://lucene.apache.org/

[17] S. Helmut, Improvements in Part-of-Speech Tagging with an Application to German. Proceedings of the ACL SIGDAT-Workshop. Dublin, Ireland. (1995)

[18] D. Bourigault, C. Fabre, C. Frérot, M-P Jacques, S. Ozdowska. Syntex, analyseur syntaxique de corpus. Atala. Actes des 12`emes journ´ees sur le Traitement Automatique des Langues Naturelles, 2005, Dourdan, France. (2005)