Using WordNet for Concept-Based Document Indexing in Information Retrieval

Fatiha Boubekeur
Department of Computer Sciences,
Mouloud Mammeri University of Tizi-Ouzou
Algeria
amirouchefatiha@mail.ummto.dz

Mohand Boughanem, Lynda Tamine, Mariam Daoud
IRIT-SIG,
Paul Sabatier University of Toulouse III,
France
{boughane, tamine, daoud }@irit.fr

***Abstract*—Concept-based document indexing deals with representing documents by means of semantic entities, the concepts, rather than lexical entities, the keywords. In this paper we propose an approach for concept-based document representation and weighting. Particularly, we propose (1) an approach for concept-identification (2) and a novel concept weighting scheme. The concepts are first extracted from WordNet and then weighted by means of a new measure of their importance in the document. Our conceptual indexing approach outperforms better than classical keyword-based approaches, and preliminary tests with the weighting scheme give better results than the classical tf-idf approach.**

Keywords-Information retrieval; conceptual indexing; concept weighting; WordNet.

I. INTRODUCTION

Information retrieval (IR) is concerned with selecting from a collection of documents those that are likely to be relevant to a user information need expressed using a query. Two basic functions are carried out in an information retrieval system (IRS): document indexing and query-document matching. The main objective of indexing is to assign to each document (respectively query) a descriptor represented with a set of features, usually weighted keywords, derived from the document (respectively query) content. The main goal of query-document matching, also called query evaluation, is to estimate the relevance of a document with respect to the query. A key characteristic of classical IR models is that the degree of query-document matching depends on the number of the shared terms. This leads to critical problems induced by disparity and ambiguity.

- Disparity refers to the property that has some terms to be represented by different words and associated to identical or related senses. Disparity causes relevant document to not be retrieved. For example, a document on *unix*, nevertheless relevant for a query on *operating systems*, will not be retrieved if the words *operating* and *systems* are absent in this document.
- Ambiguity refers to two properties: homonymy and polysemy [14]. Homonymy refers the property that has some terms, represented by the same word, to be associated to different meanings. The *bark* of a dog versus the *bark* of a tree is an example of homonymy. Polysemy is related to the property of some words to express different meanings. *Opening* a door versus *opening* a book is an example of polysemy. In classic IRS, ambiguity causes irrelevant documents to be retrieved. For example, a document on *politics in France*, nevertheless not relevant for a query on *Anatole France*, will be retrieved because of the shared word *France*. Various approaches and techniques have attempted to tackle these problems by enhancing the document representation or query formulation. Attempts in document representation improvements are related to the use of semantics in the indexing process. Semantic indexing aims at representing documents (and queries) by means of senses (concepts) rather than simple words. Senses are identified (ie. disambiguated) by means of *word sense disambiguation* (WSD) approaches that allow finding the right sense of a word in a given context. WSD are classified in supervised and unsupervised approaches [32]:
- Supervised WSD uses training Corpora [8][15][19] to first build the required knowledge base for disambiguating senses. The related approach consists on examining a number of contexts of the target word (that is the word to disambiguate), in a training corpus, from which rules on word arrangement (co-occurrence, ordering, contiguity) [29], or word usage [24] are constructed. This knowledge is then used for further recognition of word sense in a given context.
- Unsupervised WSD use external linguistic resources such as MRD (*Machine Readable Dictionnary*) [11] [16][27][30], thesaurus [31], ontologies [21][25] or Wikipedia [18] in order to identify word senses instead of using "trained" senses. This is called conceptual (or concept-based) indexing.

In this paper, we propose a conceptual indexing approach based on the use of a linguistic resource namely WordNet. The main idea of our approach is to classify document words into WordNet entries, then to associate them with correct senses. We propose to use WordNet [20] as source of evidence for word sense identification and for sense weighting.

The paper is structured as follows: Section II introduces the problems of semantic indexing and then reports some related works and presents our motivations. In Section III, we detail our proposed semantic indexing approach.

 ISBN: 978-1-61208-104-5

Preliminary experimental results are presented in Section IV. Section V concludes the paper.

II. RELATED WORK AND MOTIVATIONS

A. The Problem

Conceptual indexing approaches generally rely on deriving concepts from linguistic resources such as MRD, thesaurus, and ontologies in order to identify the relevant sense (concept) of a word in a given context. For this aim, the indexing process poses two key problems: concept identification and concept weighting.

- Concept identification aims at assigning mono-words or multi-words to the most accurate entries in the ontology. Identifying representative words is a classical indexing problem. Classical approaches are based on linguistic (tokenization, lemmatization, stop-words eliminating) and statistical techniques to identify keywords in the document. Given these keywords, a key problem in semantic indexing is to identify for each keyword its right sense(s) in the document. This leads to a WSD problem.
- Concept weighting. The purpose of concept weighting is to quantify the degree of importance of each concept in the document. Weighting is a crucial problem in IR. Indeed, the quality of retrieving depends on the quality of weighting. Good weighting is required to guarantee that the relevant documents are retrieved for a given query. In classical IRS, the well known *tf*idf* weighting scheme is successfully used. In the context of conceptual indexing, the challenge is how to correctly weight concepts.

In what follows, we give an overview of the WordNet structure, a survey of related works and then highlight the key points of our approach.

B. WordNet Overview

WordNet is an electronic lexical database [20] which covers the majority of names, verbs, adjectives and adverbs of the English language, which are structured in a network of nodes and links.

1) Nodes: also called synsets are sets of synonyms.

- A synset is a concept.
- A concept, which is a semantic entity, is lexically represented by a term.
- A term is a word (mono-word term) or a group of words (multi-word term) that represents a concept.

2) Links: Links represent semantic relations between synsets, in which the hypernym-hyponym relations defined as follows:

- the *is-a* relation (also called *subsumption* relation) associates a general concept (the hypernym) to a more specific one (its hyponym). For example, the name *tower*#1[1] has as hyponyms *silo*, *minaret*, *pylo*… The *is-a* relation thus organizes WordNet synsets into a hierarchy of concepts. An example of hierarchy of synsets corresponding to the word "*dog*" is given in Table 1.
- the *instance* relation links a concept (hypernym) with its instance (hyponym). For example, the name *tower*#1 has for instance "*Eiffel tower*".

TABLE I. WORDNET SYNSETS OF THE WORD "DOG"

Noun
S: (n) dog, domestic dog, Canis familiaris (a member of the genus Canis (probably descended from the common wolf) that has been domesticated by man since prehistoric times; occurs in many breeds) "the dog barked all night"
S: (n) frump, dog (a dull unattractive unpleasant girl or woman) "she got a reputation as a frump"; "she's a real dog"
S: (n) dog (informal term for a man) "you lucky dog"
S: (n) cad, bounder, blackguard, dog, hound, heel (someone who is morally reprehensible) "you dirty dog"
S: (n) frank, frankfurter, hotdog, hot dog, dog, wiener, wienerwurst, weenie (a smooth-textured sausage of minced beef or pork usually smoked; often served on a bread roll)
S: (n) pawl, detent, click, dog (a hinged catch that fits into a notch of a ratchet to move a wheel forward or prevent it from moving backward)
S: (n) andiron, firedog, dog, dog-iron (metal supports for logs in a fireplace) "the andirons were too hot to touch"
Verb
S: (v) chase, chase after, trail, tail, tag, give chase, dog, go after, track (go after with the intent to catch) "The policeman chased the mugger down the alley"; "the dog chased the rabbit"

C. Related Work

Conceptual indexing approaches represent documents by concepts. These concepts are extracted from ontologies and other linguistic resources. The indexing process generally runs in three steps: (1) keyword extraction, (2) sense identification and (3) concept weighting.

1) Keyword extraction: keywords are extracted from the document by a classical indexing approach (tokenisation, elimination of empty words, then lemmatization) [1][2][3][4][13][26][28]. Keywords are then mapped on the ontology in order to identify the corresponding concepts (or sense). As an ambiguous term may correspond to several entries (sense) in the ontology, it is must be disambiguated. To disambiguate a word sense, Voorhees [28], classifies every synset of this word on the basis of the number of words collocated between a neighborhood of this synset and the local context (the sentence in which the word occurs) of the corresponding ambiguous word. The best classified synset is then considered as the adequate sense of the ambiguous word. In a similar approach, Katz et al [26] define the local context of a word as the ordered list of words starting from the closest useful word to the left or right neighborhood until the target word. To disambiguate word sense, Katz et al. first extract words (called *selectors*) from the local context of the target word. Then the set *S* of selectors is compared with the synsets of WordNet. The synset that has the maximum words in common with *S* is selected as the adequate sense of the target word. To

[1] *tower#1* refers to the first sense of the word *tower* in wordNet.

disambiguate an ambiguous word, Khan et al. [13], proposed an approach based on the semantic closeness of concepts. The semantic closeness of two concepts is calculated by a score based on their mutual minimal distance in the ontology. The concepts that have the highest scores are then selected. Based on the principle that, among the various possible senses (candidate senses) of a word, the most adequate one maximises its relations with other document word candidate senses, Baziz et al. [1], assign a score to every candidate sense (candidate concept) of a given word in the document. The score of a candidate concept is obtained by adding its semantic relatedness values [16][17][22] with other candidate concepts in the document. The candidate concept having the highest score is then selected as the adequate sense (concept-sense) of the associated index word. In our approach proposed in [3][4] this score is based on the sum of its similarity value with other candidate concepts in the document, balanced by their respective frequencies.

2) Concept weighting: Analogously to term weighting in classical keyword-based IRS, weighting concepts aims at assigning to each concept its importance in a document. Weighting concepts approaches decline in two main tendencies: (1) lexical weighting (2) semantic weighting approaches.

In lexical weighting approaches, the lexical concept weighting, concepts are considered through the terms which represent them. Hence, concept weighting consists on term weighting. The weighting approaches of Baziz et al [1] and Voorhees [28] relie on this principle. Based on the extended vectorial model introduced in [10], in which every vector consists of a set of sub-vectors of various concept types (called *ctypes*), Voorhees [28] proposed to weight concepts by using a normalized classic *tf*idf* scheme. The approach proposed by Baziz et al. [1], extends the *tf*idf* scheme to take into account compound terms. The proposed approach, called *Cf*idf*, allows to weight simple terms and compound terms associated with concepts. Indeed, the weight of a term is based on the cumulative frequency of the term itself and of its components.

While in the semantic concept weighting approaches, concepts are considered through their senses. Concept weighting approaches aim at evaluating the importance of the senses in a document content. This importance is estimated through the number of semantic relations the concept has with other concepts in a document. The approaches proposed in [5][9][12] are based on this principle. In addition to the concept weighting, semantic relations are also weighted in [12]. In the same context, in the approach proposed by Boughanem et al. [5], the number of relations of a concept with the other concepts in the document defines a measure called centrality of the concept. The authors combine centrality and specificity to estimate the importance of the concepts of a document. The specificity of a concept is its depth in the WordNet hierarchy. In our work introduced in [3][4], we focused on combining both semantic and lexical concept weighting. Indeed, we propose to weight compound terms (representing concepts) on the basis of a probabilistic measure of senses relatedness between terms and associated sub-terms and sur-terms. Practically, the weight of a given term *t* is based on a probabilistic measure of the possible senses of term *t* (noted *Sens(t)*) relatively to the senses of its sub-terms (*Sub(t)*) and its sur-terms (*Sur(t)*) [3] taking into account their respective frequencies in the document. The probability that a term *t* is a possible sense of a term *t'* is measured as the fraction of the number of *t's* senses including term *t'*, over the number of all senses of term *t*.

Formally:

$$P(t \in Sens(t')) = \frac{|\{C \in Sens(t') / t \in C\}|}{|Sens(t')|}. \quad (1)$$

$$W_{t,d} = \begin{pmatrix} tf(t) \\ +\sum_i tf(Sur_i(t)) \\ +\sum_j \left[P(t \in S(Sub_j(t))) * tf(Sub_j(t))\right] \end{pmatrix} * \ln\left(\frac{N}{df(t)}\right). \quad (2)$$

Where *N* represents the total number of documents in the corpus and *df(t)* the document frequency.

D. Our Contribution

Our approach proposed in this paper is a revisited version of the theoretic framework proposed in [3][4]. The key objective of our approach is to represent the document by a semantic kernel, composed of weighted concepts extracted from WordNet. In this paper, we redefine the approach of concept identification and concept weighting as follows:

1) The proposed approach of concept identification in this paper is based on the overlapping degree between a WordNet synset and the local context (the sentence) in which the word appears in the document. Unlike the approach proposed in [3], this approach presents the advantage to allow the detection of collocation of words independently of their order of appearance in the context.

2) The weighting approach proposed in this paper is based on a new measure of concept importance in a document. This measure takes into account semantic relatedness between concepts on one hand, and the concept frequency in the document on the other hand. The concept frequency is revisited so as to take into account multi-word representations of a concept.

III. OUR CONCEPTUAL DOCUMENT INDEXING APPROACH

We propose to use WordNet to build the document representative semantic index. The document indexing process is handled through three main steps: (1) Identifying WordNet concepts, (2) Assigning concepts to document

 ISBN: 978-1-61208-104-5

index terms and (3) Weighting concepts. In the following, we present these steps.

A. Concept Identification

The purpose of this step is to identify WordNet concepts that correspond to document words. Concept identification is based on the overlap of the local context of the analyzed word with every corresponding WordNet entry. The entry which maximizes the overlap is selected as a possible sense of the analyzed word. The concept identification algorithm is given in Table 2.

TABLE II. CONCEPT IDENTIFICATION ALGORITHM

Input: document *d*

Output: $N(d)$, the set of all WordNet concepts belonging to terms (words or word- collocations) in *d.*

Procedure:

Let w_i be the next word (assumed not to be a stop word), to analyze in the document *d*. We define ψ_i the context of word in the document as the sentence in *d* that contains the word occurrence being analyzed:

1. Compute $\zeta_i = \{C_1, C_2, ..., C_n\}$ the of WordNet entries containing w_i. Each $C_j \in \zeta_i$ is represented by a multi-word or mono-word term.
2. ζ_i is ranked as follows: $\zeta_i = \{C_{(1)}, C_{(2)}, ..., C_{(n)}\}$ where $j = (1), ..., (n)$ is an index permutation such as $|C_{(1)}| \geq |C_{(2)}| \geq ... \geq |C_{(n)}|$, where $|\ |$ denotes the concept length, in terms of number of words in the corresponding terms..
3. For each element $C_{(j)}$ in ζ_i, do:

- Compute the intersection $\eta = \cap(\psi_i, C_{(j)})$ as the set of common words between ψ_i and the representative term of C_j.
- If $|\eta| < |C_{(j)}|$ then the concept-sense $C_{(j)}$ is not within the context ψ_i
- If $|\eta| = |C_{(j)}|$ then the concept-sense $C_{(j)}$ is within the context ψ_i. $C_{(j)}$ is added to the set of possible senses associated with the document

4. The process is repeated for each concept sense C in ζ_i, for which $|C| = |C_{(j)}|$.

B. Term Disambiguation

Each term t_i in document d may be associated to a number of related possible senses ("i.e." WordNet concepts) S_i. To disambiguate a term t_i, we associate a score to each of its possible senses, based on its semantic relatedness to other concepts in *N(d)*. The concept C_i which maximizes the score is then selected as the best sense of term t_i.

Formally:

$$C_i = \arg\max_{C_i \in S_i} \left(\sum_{\substack{\grave{a} \leq j \leq |N(d)| \\ j \neq i}} \sum_{c_k \in S_j} occ(C_i) * occ(C_k) * Dist(C_i, C_k) \right) \quad (3)$$

Where $occ(C_i)$ is the number of C_i's occurrences in the document, and $Dist(C_i, C_k)$ is the semantic relatedness between concepts C_i and C_k.

The set of all selected senses represents the semantic core of the document *d*.

C. Concept Weighting

Our objective here is to assign to each concept in *N(d)*, a weight that expresses its importance. For this aim, we first introduce some definitions and then present our concept weighting approach.

1) Definitions: Let *C* and *C'* be two concepts in *N(d)*. *C* and *C'* are represented by terms *t* and *t'* respectively.

Definition 1: *t'* is a sub-term of *t*, if the set of words that compose *t* includes the set of words that compose *t'*.

Definition 2: *C'* is a sub-concept of *C*, if *t'* is a sub-term of *t*.

Let $Sub_j(C)$ be the set of all sub-concepts of concept *C*. We note $Sens(C)$ the set of all WordNet senses semantically related to *C*.

Definition 3: *C'* is a possible sense of *C*, if $C' \in Sens(C)$.

2) The Weighting approach: Our concept weighting approach is based on the following assumptions:

- the more a concept is frequent and strongly correlated to other concepts in the document, the more it is important,
- The frequency of a concept relies on its occurrences and the occurrences of its sub-concepts in the document.

Based on these assumptions, we propose a concept weight scheme based on:

- The semantic relatedness, $Dist(C_i, C_j)$, between the considered concept C_i and other concepts C_j in *N(d)*.
- The frequencies of the related concepts. The frequency of a given concept *C* depends on its own occurrences in the document, and on the occurrences of its sub-concepts $Sub_j(C) \in N(d)$, balanced by the probability that the sub-concept expresses a related meaning to the concept.

Formally:

$$W(C_i) = \sum_{i \neq j, 0 \leq i, j \leq |N(d)|} tf(C_i) * tf(C_j) * Dist(C_i, C_j). \quad (4)$$

And:

$$tf(C_i) = occ(C_i) + \sum_{C_k \in Sub(C),} occ(C_k) * P(C_k \in (Sens(C_i))). \quad (5)$$

Where $P(C_k \in (Sens(C_i)))$ is the probability that C_k is a related sense of C_i.

Formally:

$$P(C_k \in (Sens(C_i))) = \frac{Dist(C_i, C_k)}{\max_{C_j \in Sens(C_i)} (Dist(C_i, C_j))}. \quad (6)$$

IV. EXPERIMENTAL EVALUATION

Our evaluation objective is to (1) measure the effectiveness of our proposed approach compared to classical indexing approaches and to (2) study the effect of concept weighting approach compared to classical term weighting.

In the following, we first present the experimental settings (the test collection and the evaluation protocol), then present and discuss the evaluation results of both our concept identification and concept weighting approaches.

1) The Test Collection: For our experiments, we used Muchmore test collection [7]. Muchmore is a parallel corpus of English-German scientific medical summaries obtained from the Web site of Springer. It declines in two versions among which an annotated one and a non annotated one. We used only the collection of non annotated English texts. This latter consists of 7823 documents and 25 queries. Relevant assessments are associated with each query.

2) Evaluation Protocol: The approach is evaluated using Mercure IR system [6]. The evaluation is made according to the TREC protocol. More precisely, every query is submitted to the system with the fixed parameters. The system returns the first 1000 documents for each query. The precision P5, P10, P20 and MAP (average precision) are computed. The precision *Px* at point *x* (*x*=5, 10, 20), is the ratio of the relevant documents among the first *x* returned documents. MAP is the mean average precision. We then compared the results obtained from our approach to different baselines.

A. Evaluation of Concept Identification Approach

Our objective of this experiment is to evaluate the impact of the semantic index quality on the retrieval effectiveness. For this aim, we compare two indexes:

- The first one is the semantic index composed of concepts, identified using our concept identification approach introduced in Section III, where each concept is weighted by means of *tf*. This approach is noted Concepts-TF in Figure 1.
- The second index is composed of a combination of both concepts and simple keywords weighted by means of *tf*. Keywords refer to those words that have no entries in WordNet. This approach is noted Concept-Fusion in Figure 1.

Retrieval results obtained using each of these two indexes are compared to two baselines:

- The first one is a classic baseline based on keyword-based indexing, where terms are weighted by means of classical *tf*idf* scheme. This approach is noted Classic-TFIDF in Figure 1.
- The second baseline is based on a keyword-based indexing where terms are weighted according to the BM25 scoring function [23].

Remark: No comparison was made with our approach proposed in [3][4], which mainly remains a theoretical framework. Indeed, this latter approch was not fully implemented (due to the complexity of its induced calculations), and only partial related results were available.

The evaluation results obtained for these different models are presented in Figure 1. According to the results, we conclude the following:

- Concepts-TF approach is better than the Classic-TFIDF baseline. The percentage of improvement is of 61 % for P5, 51 % for P10, 54 % for P20 and 51 % for the MAP
- The Concepts-Fusion approach is better than the Concepts-TF approach. The percentage of improvement is of 20 % for P5, 19 % for P10, 15 % for P20 and 23 % for the MAP. To study the statistical significance of these improvements, we have calculated the Wilcoxon signed-rank test between each indexing model and the baseline search performed by *tf*idf* weighting scheme. We assume that the difference between models is significant if the p-value p <0.1 and very significant if p<0.05. We have obtained a very significant p-value according to the Wilcoxon test of our model compared to classical indexing at almost the precision, P5, P10, P20 and MAP (see Table III). This proves the statistical significance of our indexing model to classical one. These results consolidate us in the idea that a combined indexing concepts+keywords is more effective than a concept-based indexing.

TABLE III. STATISTICAL RESULTS FROM WILCOXON TEST

p-value at	Classic-TFIDF vs. Concept-TF P	Classic-TFIDF vs. Concept-Fusion P
P5	0,0015	< 0,0001
P10	0,0081	0,0002
P20	0,0042	0,0001
MAP	0,0102	< 0,0001

- Besides, our Concepts-Fusion approach presents better results than Classic-TF baseline with increasing rates of 94 % for P5, 45 % for P10, 77 % for P20 and 77 % for the MAP. Nevertheless, as

shown on Figure 1, the Concepts-Fusion approach results are worse than those of the Classic-OKAPI baseline with decreasing rates of 0 % for P5, -1 % for P10, -5 % for P20 and -3 % for the MAP. This shortcoming is probably due to the imprecision of the disambiguation approach. Indeed, in a context of a precise disambiguation, we expect that indexing by the concepts will bring higher performance than indexing with keywords.

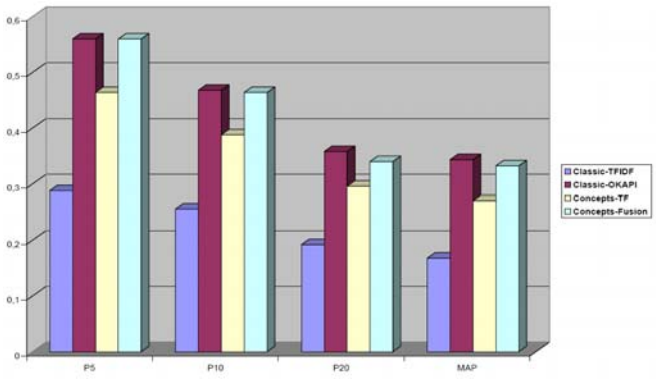

Figure 1. Concept vs keyword indexing.

B. *Evaluation of Concept Weighting Approach*

The second series of our experiments focuses on the evaluation of our concept-weighting approach introduced in Section II.C. Practically, we aim at measuring the impact of the weighting-scheme on the retrieval effectiveness. For this aim, we compare the effectiveness of two indexes:

- The first one consists on the concepts detected by our approach proposed in Section II.B, balanced by their respective frequencies. This approach is noted Concepts-TF in Figure 2.
- The second index consists on the concepts detected by our approach proposed in Section II.B, balanced by the proposed weight defined in Section II.C. This approach is noted Concepts-Score in Figure 2.

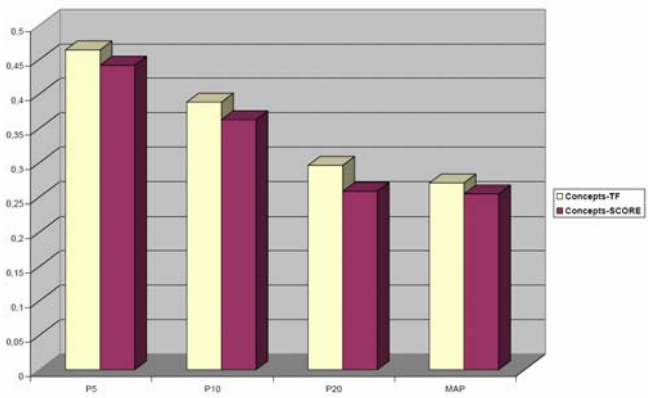

Figure 2. TF vs Score weighting in concept-based indexing.

Figure 2 presents a comparison between these two weighting approaches. From this figure, it appears that the results obtained from our proposed concept-weighting approach are globally less effective compared to those obtained from the frequency-based concept weighting scheme, with decreasing rates of -5 % for P5, -6 % for P10, -12 % for P20 and -6 % for the MAP. The obtained results are clearly below of our expectations. The problem behind this shortcoming improvement is probably due to the ranking score, used by Mercure search engine [6] to estimate the correspondence of a document to a query. Indeed, in evaluating the Concepts-Score index, instead of a *tf-idf* combination, the ranking score combines the concept weight with the non-correlated *idf* measure. This leads to decrease the precision improvement of the retrieved results.

V. CONCLUSION AND FUTURE WORK

We have presented in this paper a novel approach for conceptual document indexing. Our contribution concerns two main aspects. The first one consists on a concept-indexing approach based on the use of WordNet. The approach is not new but we proposed new techniques to identify concepts and to weight them. Preliminary results showed that our proposed concept-identification approach is more effective than a classical keyword-based indexing approach, and brings significant increasing rates compared to the Classic-TFIDF approach. However, this approach, even if combined with keywords, does not perform as well as the Classic-OKAPI baseline, probably due to the slight imprecision of our disambiguation. Besides, the concept-weighting approach produced reserved results. The likely cause of this unexpected shortcoming is the non-relevance of the ranking score for the semantic index. In future works, we plan first to revisit our concept disambiguation approach, and second to propose a ranking score for semantic indexes, which takes into account semantic weights of concepts. Works in this direction are in progress. .

ACKNOWLEDGMENTS

This work was made possible thanks to the financial support of the A.U.F. (*Agence Universitaire de la Francophonie*) and of U.M.M.T.O. (*Université Mouloud Mammeri de Tizi-Ouzou*) with the kind collaboration of IRIT (*Institut de Recherche en Informatique de Toulouse*).

REFERENCES

[1] M. Baziz, M. Boughanem, and N. Aussenac-Gilles, “A Conceptual Indexing Approach based on Document Content Representation”, Dans: CoLIS5: Fifth International Conference on Conceptions of Libraries and Information Science, Glasgow, UK, 4 juin 8 juin 2005., F. Crestani, I. Ruthven (Eds.), Lecture Notes in Computer Science LNCS Volume 3507/2005, Springer-Verlag, Berlin Heidelberg, pp. 171-186.

[2] M. Baziz, M. Boughanem, and N. Aussenac-Gilles, “The Use of Ontology for Semantic Representation of Documents“, Dans: The 2nd Semantic Web and Information Retrieval Workshop(SWIR), SIGIR 2004, Sheffield UK, 29 juillet 2004, Ying Ding, Keith van Rijsbergen, Iad Ounis, Joemon Jose (Eds.), pp. 38-45.

[3] F. Boubekeur, M. Boughanem, and L. Tamine, “Exploiting association rules and ontology for semantic document indexing“, Dans: 12th International conference IPMU08, Information Processing and Management of Uncertainty in knowledge-Based Systems, Malaga, 22- 27, June 08, Spain, pp. 464-472.

 ISBN: 978-1-61208-104-5

[4] F. Boubekeur, M. Boughanem, and L.Tamine, “Semantic Information Retrieval Based on CP-Nets”, Dans: IEEE International Conference on Fuzzy Systems (FUZZ-IEEE 2007), London, 23/07/07- 26/07/07, IEEE, July 2007, pp. 1-7.

[5] M. Boughanem, I. Mallak, and H. Prade, “A new factor for computing the relevance of a document to a query”, Dans: IEEE World Congress on Computational Intelligence (WCCI 2010), Barcelone, July 2010, (to appear).

[6] M. Boughanem and C. Soulé-Dupuy, “A Connexionist Model for Information Retrieval“, DEXA 1992, pp. 260-265.

[7] P. Buitelaar and H. Uszkoreit, “MuchMore: Concept-Based Cross-Lingual Information Retrieval in the Medical Domain”, In: Kuenstliche Intelligenz, Heft 2/04, 2004, pp. 43-44, (http://muchmore.dfki.de/resources1.htm, 2010).

[8] M. Cuadros, JM., Atserias, J., M. Castillo, M., and G. Rigau, “Automatic acquisition of sense examples using exretriever“, In *IBERAMIA Workshop on Lexical Resources and The Web for Word Sense Disambiguation*. Puebla, Mexico, 2004, pp. 97-104.

[9] D. Dinh and L. Tamine, “Vers un modèle d'indexation sémantique adapté aux dossiers médicaux de patients“, Dans: Conférence francophone en Recherche d'Information et Applications (CORIA 2010), Sousse, Tunisia, March 2010, Hermès editions, (electronic support).

[10] E.A. Fox, “Extending the boolean and vector space models of information retrieval with p-norm queries and multiple concept types”, PhD thesis, Ithaca, NY, USA, 1983.

[11] J.A Guthrie, L. Guthrie, Y. Wilks, and H. Aidinejad, “Subject-dependant cooccurrence and word sense disambiguation“, In Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics, Berkley, CA, pp. 146-152.

[12] B.Y. Kang and S.J. Lee, “Document indexing: a concept-based approach to term weight estimation“, In Journal of Information Processing & Management. Volume 41, Issue 5, September 2005, pp. 1065-1080.

[13] L.R. Khan, D. Mc Leod, and E.Hovy, “Retrieval effectiveness of an ontology-based model for information selection“, The VLDB Journal (2004)13, pp. 71–85.

[14] R. Krovetz, “Homonymy and polysemy in information retrieval”, In Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics (ACL-97}, pp. 72-79.

[15] C. Leacock, G.A. Miller, and M. Chodorow, “Using corpus statistics and WordNet relations for sense identification“, Comput. Linguist. 24, 1 (March 1998), pp. 147-165.

[16] M.E. Lesk, “Automatic sense disambiguation using machine readable dictionaries: How to tell a pine cone from a nice cream cone”, In Proceedings of the SIGDOC Conference. Toronto, 1986, pp. 24-26.

[17] D. Lin, “An information-theoretic definition of similarity“, In Proceedings of 15th International Conference On Machine Learning, 1998, pp. 296-304.

[18] O. Medelyan, D. Milne, C. Legg and I.H. Witten, ”Mining meaning from Wikipedia“, In International Journal of Human-Computer Studies archive, Volume 67, Issue 9, September 2009, pp. 716-754, ISSN: 1071-5819.

[19] R. Mihalcea and D. Moldovan, “Semantic indexing using WordNet senses”, In Proceedings of ACL Workshop on IR & NLP, Hong Kong, October 2000, pp. 35-45.

[20] G. Miller, “WordNet: A Lexical database for English”, Actes de ACM 38, pp. 39-41.

[21] P. Resnik, “Disambiguating noun groupings with repect to WordNet senses”,, 3thWorkshop on Very Large Corpora, 1995, pp. 54–68.

[22] 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 (JAIR), 11, 1999, pp. 95-130.

[23] S.E. Robertson, “The probability ranking principle in IR”, Journal of Documentation 33, 1977, pp. 294-304. Reprinted in: K. Sparck Jones and P. Willett (eds), Readings in Information Retrieval. Morgan Kaufmann, 1997, pp. 281-286.

[24] H. Schütze and J. Pedersen, “Information retrieval based on word senses”, In Proceedings of the 4th Annual Symposium on Document Analysis and Information Retrieval, pp. 161-175.

[25] M. Sussna, “Word sense disambiguation for free-text indexing using a massive semantic network”, 2nd International Conference on Information and Knowledge Management (CIKM-1993), pp. 67–74.

[26] O. Uzuner, B. Katz, and D. Yuret, “Word Sense Disambiguation for Information Retrieval”, AAAI/IAAI 1999, pp. 985.

[27] J. Véronis and N. Ide., “Word sense disambiguation with very large neural networks extracted from machine readable dictionaries”, In 13th International Conference on Computational Linguistics (COLING-1990), 2, 1990, pp. 389-394.

[28] E. M. Voorhees, “Using WordNet to disambiguate word senses for text retrieval”, Association for Computing Machinery Special Interest Group on Information Retrieval. (ACM-SIGIR-1993): 16thAnnual International Conference on Research and Development in Information Retrieval, 1993, pp. 171-180.

[29] S.F. Weiss, “Learning to disambiguate”, In Information Storage and Retrieval, 9, 1973, pp. 33-41.

[30] Y. Wilks and M. Stevenson, “Combining independent knowledge source for word sense disambiguation”, Conference «Recent Advances in Natural Language Processing », 1997, pp. 1-7.

[31] D. Yarowsky., “Word-Sense Disambiguation Using Statistical Models of Roget's Categories Trained on Large Corpora”, In Proceedings of the 14th International Conference on Computational Linguistics (COLING-92). Nantes, France, August 1992, pp. 454-460.

[32] D. Yarowsky, “Unsupervised word sense disambiguation rivaling supervised methods”, In 33rd Annual Meeting, Association for Computational Linguistics, Cambridge, Massachusetts, USA, 1995, pp. 189-196.

 ISBN: 978-1-61208-104-5