World Academy of Science, Engineering and Technology
International Journal of Computer and Information Engineering
Vol:1, No:1, 2007

Concept Indexing using Ontology and Supervised Machine Learning

Rossitza M. Setchi, and Qiao Tang

Abstract**—**Nowadays, ontologies are the only widely accepted paradigm for the management of sharable and reusable knowledge in a way that allows its automatic interpretation. They are collaboratively created across the Web and used to index, search and annotate documents. The vast majority of the ontology based approaches, however, focus on indexing texts at document level. Recently, with the advances in ontological engineering, it became clear that information indexing can largely benefit from the use of general purpose ontologies which aid the indexing of documents at word level. This paper presents a concept indexing algorithm, which adds ontology information to words and phrases and allows full text to be searched, browsed and analyzed at different levels of abstraction. This algorithm uses a general purpose ontology, OntoRo, and an ontologically tagged corpus, OntoCorp, both developed for the purpose of this research. OntoRo and OntoCorp are used in a two-stage supervised machine learning process aimed at generating ontology tagging rules. The first experimental tests show a tagging accuracy of 78.91% which is encouraging in terms of the further improvement of the algorithm.

Keywords**—**Concepts, indexing, machine learning, ontology, tagging.

I. Introduction

RECENT advances in semantic technologies made possible traditional ways of indexing to be revisited, and a number of advanced semantic-based approaches to indexing were developed. Semantic document indexing is a way of coding digital texts and images to represent their associated abstract meaning [1, 2]. Practical semantic indexing is impossible without some particular knowledge modeling commitments, as the ability of an information retrieval (IR) system to understand concepts and ideas is limited by the underlying representation system used [3].

As in most cases, the suitable representation scheme depends on the user needs. These are thoroughly explored in [1] where several scenarios of use outline the type of search a semantic-based system should be able to perform. As pointed out in this study, scholars should be provided with many different ways of conceptualizing and exploring their subjects.

Manuscript received November 17, 2006. The research described in this paper was conducted within the TRENDS project sponsored by FP6 of the European Commission.

R. M. Setchi is with Cardiff School of Engineering, Cardiff University, Cardiff CF24 3AA (phone.: +44-(0)-292087-5720; fax: +44-(0)-292087-4716; e-mail: setchi@cf.ac.uk).

Q. Tang is also with Cardiff School of Engineering, Cardiff University, Cardiff CF24 3AA (e-mail: tangq@cf.ac.uk).

Therefore, there is a need for a new generation of search engines which can provide alternative ways into the information databases tailored to the specific kinds of research.

Another study of user information retrieval needs [4] was conducted within the TRENDS research project where leading designers of concept cars were interviewed about their sources of inspiration. The study showed that these professionals require intranet and internet search engines which enable focused search for design specific elements (such as shapes, volumes, colors, and fabrics), enable categorization of information (for example, grouping cars by designer name, period of time, or market segment), and help in illustrating subjective emotions or concepts (for instance, strength, shock, danger, or fluidity).

These examples clearly show the need for semantic-based representation which uses both *instances* and *abstract ideas*. This conclusion is indirectly supported by the research reported in [5] which claims that "as far as the human brain is concerned, it is unrealistic to treat a keyword as the sole representation of a concept."

This paper supports the view that document indexing based on identifying entities (or instances) and concepts (abstract ideas) in a document supports different levels of abstraction and different information retrieval needs. The paper describes a semantic based indexing approach which links words and phrases to a general purpose ontology *OntoRo*. The algorithm uses supervised machine learning which benefits from the use of another linguistic resource, *OntoCorp*, specifically developed for the purpose of this research.

The paper is organized as follows. Section II introduces the use of ontologies in semantic-based indexing. Section III describes the proposed concept indexing framework. Section IV describes the ontology tagging algorithm developed, and the tests conducted to prove the feasibility of the approach. Finally, section V presents conclusions and directions for further research.

II. Ontology-Based Approaches to Semantic Indexing

An ontology specifies a conceptualization of a domain in terms of concepts, attributes and relations. Concepts are typically organized into a tree structure; in addition, they are linked through relations forming a semantic net structure.

Nowadays, ontologies are the only widely accepted paradigm for the management of open, sharable, and reusable knowledge in a way, which allows automatic interpretation [3,

6]. They provide background knowledge, views and navigation structures for browsing. They support integration of knowledge sources as they build upon a collective understanding within a community. Today, many ontologies are collaboratively created across the Web and used to search and annotate documents. Besides large well-known lexical ontologies such as the WordNet [7], there are many purpose-built ontologies. For example, the KIM system for semantic annotation, indexing and retrieval [3] uses an upper-level ontology comprising some general philosophical categories and the most common entity types (people, cities, companies, etc.). It consists of about 250 classes and 100 properties. Domain-specific dictionaries and ontologies are also used to improve tagging and ultimately IR [8]. Medical and life sciences domains are typical examples where the use of domain-specific ontologies produces good results.

It needs to be noted that the vast majority of the ontology based approaches focus on indexing texts at *document* level. Recently, however, with the advances made in ontological engineering, it became clear that information retrieval and concept indexing in particular can largely benefit from the use of ontologies to index documents at *word level* [9]. For example, a recent study reported in [10] uses Natural Language Processing (NLP) technology and three ontologies (WordNet, OpenCyc and SUMO) to index texts *word by word*. These three ontologies contain 4115 concepts. The mapping accuracy between the three ontologies is 96.2% while the accuracy of the ontology tagging is estimated to be between 60% and 70% [10].

The approach described in this paper supports part of speech (POS) tagging, word sense disambiguation and the retrieval of texts that contain similar words by indexing them to concepts contained in an ontology.

III. Concept Indexing Framework

A. Concept Indexing

In the context of this paper, *concept indexing* is defined as the process of identifying instances (entities) and abstract ideas (concepts) within a text document, and linking the words and phrases in a text to ontological concepts. *Concept index* is a machine understandable index of entities and concepts contained in document collections. An *entity* is an identifiable and discrete instance existing in a text document. A *concept* is an abstract or general idea inferred or derived from specific instances.

The main assumption underlying concept indexing is that the information conveyed in a text can be analyzed in terms of the entities and concepts that text contains. This approach involves three steps:

(i) extracting entities from unstructured text-based content using lexical tags and rules,

(ii) identifying concepts and adding ontology tags to them using semantic rules, and

(iii) merging entity and concept information into a concept index.

For example, in the following sentence from the Brown Corpus [11],

The Fulton County Grand Jury said Friday an investigation of Atlanta's recent primary election produced no evidence that any irregularities took place.

(1)

the *entity information* is

Fulton/NP County/NN, Jury/NN, Friday/NR, investigation/NN, primary/NN election/NN, evidence/NN, irregularities/NNS, place/NN

(2)

and the *concept information* is

grand/S536 jury/S280 say/S312 friday/S65 investigation/S267 recent/S75 primary/S313 election/S362 produce/S470 evidence/S322 irregularity/S48

(3)

The tags attached to the entities are the *Brown POS tags* employed in the tagged version of the Brown Corpus [11]. In it, each individual word is given a grammatical tag from a list of 81, each specifying a particular word-class. For example, "NP" means proper noun or name phrase while "NNS" indicates a plural noun.

The concept information contains ontology tags (called in this research *OntoRo tags*) indicating a concept group within the ontology OntoRo used in this research. For instance, S267 is the ontology tag attached to the word "investigation". This concept group (#267) contains 672 different entries; some of these entries are synonyms (i.e. "inquiry", "questioning" and "examination") while others (i.e. challenge", "McCarthysm" and "CIA") are not synonyms but within a certain context could be semantically linked to the word "investigation".

B. Conceptual Framework

The framework in Fig. 1 shows that the concept indexing is performed by the *POS tagger* and the *ontology tagger*.

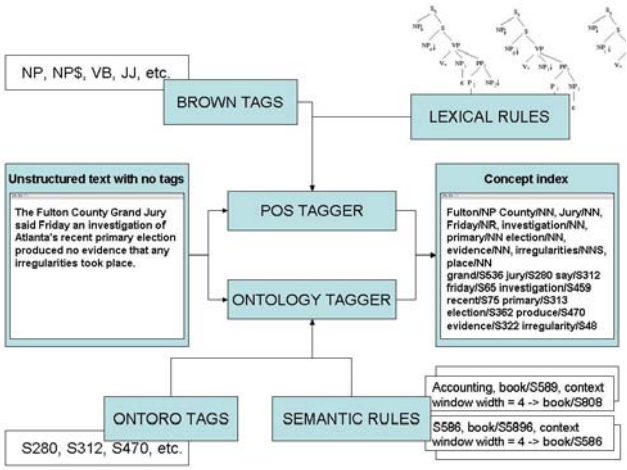

Fig. 1 Conceptual framework

The POS tagger uses *lexical rules* to attach Brown tags to all words in the text, and then identify the entities in it. For example, sentence (1) is first fully tagged as (4) and then the entity information is extracted as in (2).

the/AT Fulton/NP County/NN Grand/JJ Jury/NN said/VBD Friday/NR an/AT investigation/NN of/IN Atlanta's/NP$ recent/JJ primary/NN election/NN produced/VBD no/AT evidence/NN that/CS any/DTI irregularities/NNS took/VBD place/NN ./.

(4)

The POS tagger used in this research is based on the Brill tagger [12]. It uses 508 lexical rules similar to those shown in (5) [13].

TO IN NEXTTAG PP$
VBN VBD SURROUNDTAG CC AT
VB NN PREVTAG AP

(5)

The ontology tagger uses *semantic rules* (6) to attach ontology tags to the concepts in the text. The semantic rules are result of supervised machine learning which will be explained later.

Accounting, book/S350, context window width = 4 -> book/S496
S347, book/S350, context window width = 4 -> book/S347

(6)

The ontology used to assign ontology tags to concepts is OntoRo [14]. It was built using the printed [15] and the electronic (project Gutenberg's) [16] version of the Roget's Thesaurus (Roget's). The process of building OntoRo is described in detail in [14]. OntoRo contains 68,920 unique words and 228,130 entries. These are classified into 990 concepts, 610 head groups, 95 subsections, 39 sections, and 6 top level classes. The format of an OntoRo entry is as follows.

Concept, POS, concept group, word/phrase, head/group, subsection, section, top level

(8)

Examples of OntoRo entries are shown below.

459,n,1,investigation,267,50,17,4
459,n,1,examination,267,50,17,4
459,n,7,CIA,267,50,17,4
536,n,2,investigation,314,57,24,4
438,n,4,examination,255,48,15,3
449,n,1,examination,261,49,16,4
455,n,1,examination,265,50,17,4

(7)

The feasibility of the proposed concept indexing approach depends on the availability of semantic rules with good classification accuracy. These rules are obtained using the machine learning (ML) process described in the next section. As in all NLP applications based on ML, the availability of training and testing material, in this case a corpus annotated with OntoRo tags, is critical.

IV. Training of the Ontology Tagger

A. Ontologically Tagged Corpus

OntoCorp, the corpus with OntoRo tags, which is used as a source of training and testing material, is built using an existing standard corpus. This is *SemCor* [17], a well known and well studied corpus for semantic analysis that is supplied with WordNet 1.6. Semcor is a package of semantic concordance text annotated using information from WordNet. A semantic concordance is a textual corpus and a dictionary combined in such a way that every word/phrase in the text is linked with its appropriate sense in the dictionary. All words in Semcor are annotated using tags with attribute - value pairs.

The difficulty in creating OntoCorp lies in the different structure and organization of WordNet and OntoRo. The development of OntoCorp involves three steps:

(i) Building a machine readable dictionary *eWord* from WordNet (needed as WordNet uses a purpose built database),

(ii) Mapping the eWord (WordNet) and OntoRo entries (many-to-many mapping problem), and

(iii) Converting SemCor (tagged using WordNet senses) into OntoCorp (tagged to OntoRo concepts).

A hypothesis was made which was tested and proved in [14] that most entries in WordNet have appropriate matching entries in OntoRo. The difficulty is that every eWord entry may have several possible OntoRo meanings to choose from, although in a certain context there may be only one suitable choice. The following hypothesis is used as a heuristic to solve this problem.

For each word/phrase in the description of a given eWord entry, the corresponding OntoRo entry is the one which appears most often.

For the eWord entry, its corresponding OntoRo entry is the one which appears most as an entry for the individual words/phrases in the description of that eWord word/phrase.

(9)

This hypothesis is based on the observation that the words/phrases in the description of an eWord entry often carry very similar meaning. Thus, if ontology tags from OntoRo are attached to them, they should reveal a certain degree of similarity.

The mapping between a given eWord entry and its corresponding OntoRo entry includes first assigning corresponding ontology entries from OntoRo to each word/phrase contained in the description of that eWord entry, and then choosing a tag among those assigned to the words/phrases in the description, which best represents the meaning of the whole eWord entry. The process continues until all entries in eWord have been assigned a tag from OntoRo.

This approach is illustrated in Fig. 2. In eWord, the word/phrase E is described through a number of words/phrases (A-D), and each of them has several possible meanings in OntoRo. According to (9), meaning no.1 for A, meaning no.1 for B, meaning no.1 for C and meaning no.2 are first identified as candidates. Next, meaning no.1 is selected to represent the meaning of the word E, as it is the most frequently used one.

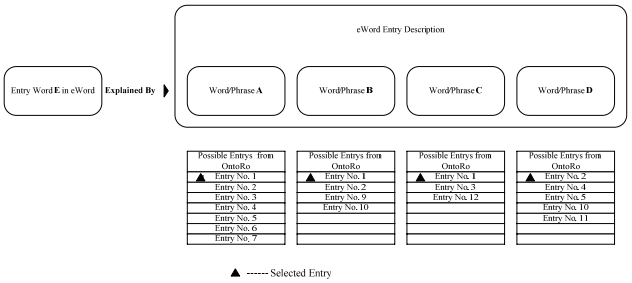

Fig. 2 "One OntoRo meaning per eWord entry" hypothesis

The mapping algorithm uses a semantic similarity measure which determines the degree to which one meaning is close to another. The semantic similarity is measured by pair-wise comparisons of the values in the fields of the ontology entries, in the following order of priority: concept, subsection, section and top level. The elements are compared on a lower priority level only if the values in the field with a higher priority are the same. If two elements have the same degree of semantic similarity, then the one that appears first in the candidate list is selected.

When semantic similarity measures cannot be applied (if there are no fields with the same values), the element with the lowest occurrence in eWord is selected from the first five words in the description of that entry. This is based on the observation that words and phrases less frequently appearing in a text carry more information [18]. The selection is limited to the first five words only, because simple tests reveal that the most appropriate OntoRo concept for a word/phrase is often among those assigned to the first five words.

Fig. 3 illustrates this algorithm by depicting possible OntoRo entries for the eWord entry shown in (10).

Abandon^v^3^vacate, empty, abandon, leave behind, move out (10)

In Fig. 3, "C" shows, for each word used in the description of an eWord entry, the number of times certain concept appears against other words from the same description. For example, the word "vacate" which is the first word used in the description of the eWord entry "abandon" (10), appears five times in OntoRo: twice in concept #190, and once in each of the following: #752, #621 and #753. Subsequently, concept #621 appears against two more words: "abandon" and "move out", hence the number 2 in the "count" field for #621. As shown in the figure, the mapping algorithm selects one possible meaning for each word in the description. Finally, the eWord entry word "abandon" is mapped to the OntoRo entry "621,1,3,371,27,62,5".

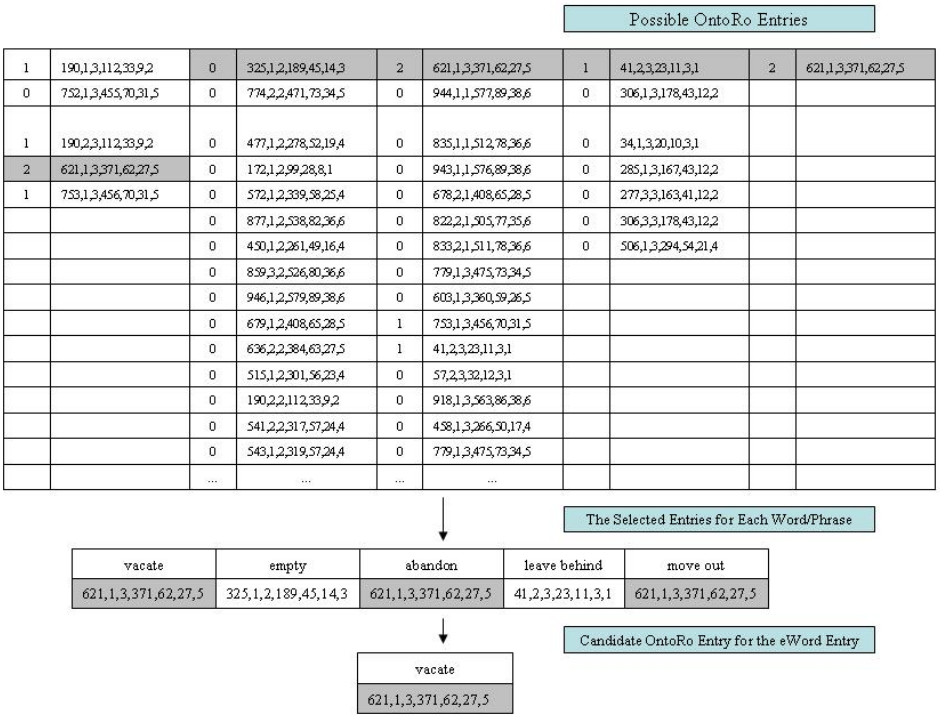

Fig. 3 Illustration of the semantic mapping algorithm

The mapping algorithm is implemented in C language. The tests were conducted on a Pentium III 700MHz 384MB memory computer. The total running time for mapping all 77,022 entries in eWord to the 37,301 OntoRo entries was 418 minutes. In addition, tests were performed to examine the mapping accuracy. 160 of the 200 samples randomly selected, were found correct according to the expert involved. The mapping accuracy is therefore 80%.

Once the semantic mapping between eWord and OntoRo is completed, the next stage of converting SemCor into OntoRo is governed by three simple rules:

1. All words/phrases with little semantic information or without "lemma" attribute in the Semcor entries, as well as all proper nouns and punctuation marks are tagged with a tag "IGNORE".

2. All words/phrases that are not found in OntoRo are tagged as "UNKNOWN".

3. The remaining words/phrases are tagged using an ontology tag composed of the corresponding concept number in the hierarchy of OntoRo proceeded by the letter "S".

A typical ontologically tagged sentence from OntoCorp is shown in (11).

His/IGNORE petition/S312 charged/S497 mental/S260 cruelty/S556 ./IGNORE

(11)

OntoCorp contains 20,138 sentences with 434,998 tagged words.

B. Semantic Rules

Two types of semantic rules are used in this work: *statistical* and *context* rules. These rules are extracted from the ontologically tagged corpus OntoCorp. An example of a statistical rule is given below.

book -> book/S350

(12)

where, "book" is a word/phrase to be tagged by the algorithm, and "S350" is a concept tag for "books and publications".

This statistical rule denotes that every word "book" in a text will be assigned a concept tag "S350" regardless of whether the concept fits the context of use or not. Therefore, statistical rules are more effective if a given word has a dominant meaning.

Statistical rules are complemented in this research by context rules like those shown in (6). In addition to concept tag "S350" mentioned above, these rules use concept "S496" related to the use of books in "accountancy and book-keeping", and "S347" associated with "writing". The first context rule in (6) means that if the word "book" has an ontology tag "S350" attached, and the word "accounting" appears within a context window of four words, then tag "S350" should be replaced by "S496". The second context rule in (6) means that if the word "book" has an ontology tag "S350" assigned and an ontology tag "S347" appears within a context window of four words around "books/S350", then tag "S350" should be changed to "S347".

C. Training and Tagging

These two types of rules are generated during the training and then utilized during the tagging process. The flowcharts shown in Fig. 4 illustrate these processes.

In the training process, first statistical and context information is obtained from OntoCorp. All tags are removed from OntoCorp to create an untagged corpus (UC). When the statistical and context rules are applied in sequence to it, the tags generated are compared with those in OntoCorp. The percentage of correctly assigned tags is used to estimate the accuracy of the algorithm.

The next step involves selecting the most frequently used ontology tag for each unique word/phrase in OntoCorp and generating statistical rules. These rules are used to tag the words/phrases in the UC. Then, the incorrectly tagged words are identified by comparing the tagged UC with OntoCorp. Next, the corresponding context information from OntoCorp for those wrongly tagged UC words/phrases is utilized to generate context rules. These context rules are then used to replace some of the attached tags and assign tags to the words/phrases which have not been tagged.

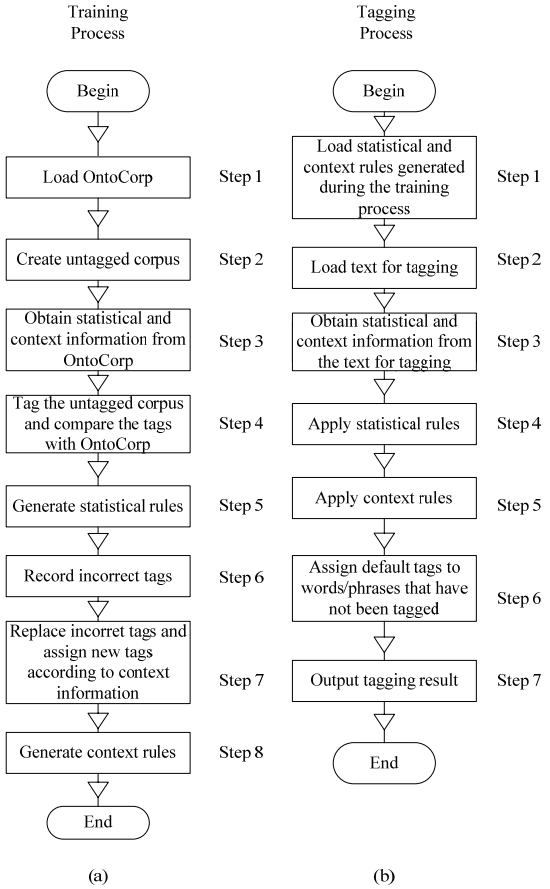

Fig. 4 Flowcharts of the ontology tagging algorithm (training and tagging)

In the tagging process, the statistical and context rules generated during the training are first loaded. Next, the text is processed and statistical information is obtained. Then, the statistical rules are used to assign ontology tags to the words in the text. After that, context rules are applied to replace some of the previously assigned tags or to assign tags to those words/phrases which have not been tagged. Finally, all untagged words/phrases are annotated with a default ontology tag.

D. Experiments

When context rules are used to replace tags, which have been previously assigned according to the statistical rules, some of the correct tags may be replaced by wrong ones. Therefore, the algorithm needs to ensure that when applying context rules to the corpus, the context rules replace more wrong tags with correct ones than correct ones with wrong ones. Due to this reason, the method of generating context rules has to be carefully designed and evaluated. This aspect is thoroughly investigated in [14] where 96 experiments were conducted to investigate the co-relation between different variables and the three possible design options for generating context rules. These are using: (i) word co-occurrence frequency, (ii) ontology tag co-occurrence frequency, and (iii) mutual information for words in a context window.

The experiments showed that the average accuracy ranges from 76.40% to 78.91%, with highest accuracy achieved when 90% of the corpus is used for training, mutual information is employed, and the context window contains 6 words. All experiments used the same method of extracting statistical rules. Therefore, the variation in the tagging accuracy is due to the method of generating context rules. Thus, the ontology tagging problem can be redefined as how to compose the context rules to be applied to the text after the tagging based on statistical rules is completed.

V. Conclusion and Future Work

This paper presents a method for concept indexing algorithm which identifies entities and concepts in a text, tags the entities using Brown tags and lexical rules, and tags the concepts with ontology tags using semantic rules. The method employs two resources developed in this research. These are the general-purpose ontology OntoRo and the ontologically tagged corpus, OntoCorp, which are used by the supervised machine learning algorithm for automatic ontology tagging developed.

Experiments were conducted to measure the mapping accuracy of eWord (the version of WordNet used) and OntoRo and the tagging accuracy of the ontology tagger.

The results show that with 90% of the corpus used for training, and using mutual information with a context window of 6 words for context rule generation, the tagging algorithm achieved 78.91% accuracy. It should be noted that the accuracy is calculated using the corpus automatically generated by mapping eWord entries into OntoRo entries, where the mapping accuracy is merely 80%. This means that the training material is not completely accurate, which brings tagging errors into the system. Therefore, the results can only be considered as an indication of the tagging accuracy. One way of improving the algorithm is by verifying all mappings between eWord and OntoRo. Future work includes further optimization of the tagging algorithm.

References

[1] S. Rhind-Tutt, "Semantic indexing: a case study", *Library Collections, Acquisitions, and Technical Services*, vol. 27, n. 2, pp. 243-248, 2003.

[2] T. Brasethvik, and J. A. Gulla, "Natural language analysis for semantic document modeling", *Data & Knowledge Engineering*, vol. 38, n. 1, pp. 45-62, 2001.

[3] A. Kiryakov, B. Popov, I. Terziev, D. Manov, and D. Ognyanoff, "Semantic annotation, indexing, and retrieval", *Web Semantics: Science, Services and Agents on the World Wide Web*, vol. 2, n. 1, pp. 49-79, 2004.

[4] TRENDS Project FP6-IST-2005-27916, "List of user specifications", http://cpn.paris.ensam.fr/trendsproject/, accessed 7 November 2006.

[5] R. K. Rajapakse, and M. Denham, "Text retrieval with more realistic concept matching and reinforcement learning", *Information Processing & Management*, vol. 42, n. 5, pp. 1260-1275, 2006.

[6] L. van Elst, and A. Abecker, "Ontologies for information management: balancing formality, stability, and sharing scope", *Expert Systems with Applications*, vol. 23, n. 4, pp. 357-366, 2002.

[7] G. A. Miller, "WORDNET: an on-line lexical database, International Journal of Lexicography, vol. 3, n. 4, 1990, pp. 235-312.

[8] A. R. Coden, S. V. Pakhomov, R. K. Ando, P. H. Duffy, and C. G. Chute, "Domain-specific language models and lexicons for tagging", *Journal of Biomedical Informatics*, vol. 38, n. 6, pp. 422-430, 2005.

[9] R. Setchi, Q. Tang, and L. Chen, "an information retrieval system using deep natural language processing", *Lecture Notes in Artificial Intelligence*, vol. 2773, pp. 879 - 885, 2003.

[10] J. Köhler, S. Philippi, M. Specht, and A. Rüegg, "Ontology based text indexing and querying for the semantic web", *Knowledge-Based Systems*, in press, available at www.sciencedirect.com 13 July 2006.

[11] W. N. Francis, H. Kucera, *Brown corpus manual of information*, to accompany *Standard Corpus of Present-Day Edited American English*, Providence, Rhode Island, Department of Linguistics, Brown University, 1964, revised 1971, revised and amplified 1979.

[12] E. Brill, "A simple rule-based part of speech tagger", Proc. *3rd Conf. on Applied NLP*, Trento, Italy, 1992, pp. 152-155.

[13] E. Brill, "Some advances in rule-based part of speech tagging", Proc. *12th National Conf. on Artificial Intelligence (AAAI-94)*, Seattle, US, 1994.

[14] Q. Tang, *Knowledge management using machine learning, NLP and ontology*, Cardiff, UK, PhD thesis, 2006.

[15] P. Roget, G. Davidson (ed.), *Thesaurus of English words and phrases*. Penguin Books, UK, 2003.

[16] Project Gutenberg: http://www.gutenberg.org, [accessed on 10 November 2006], 2006

[17] C. Fellbaum, (ed.), *WordNet: An electronic lexical database*, The MIT Press, USA, 1998.

[18] C. E. Shannon, "A Mathematical Theory of Communication", *Bell System Technical Journal*, vol. 27, 1948, pp. 379-42.