EXTENDING ONTOLOGY TREE USING NLP TECHNIQUE

BY

Sabrina T., Rosni A., T. Enyakong

Pusat Pengajian Sains Komputer,
11800 Universiti Sains Malaysia,
Pulau Pinang

Abstract: This paper proposes a method of creating a web document representation using a web ontology concepts instead of 'bag-of-words'. However, since the web domain has a very small vocabulary, we are unable to transform all or most of the keywords of the web document into web ontology concepts. This particular problem is solved by creating an extended part of the web ontology with words obtained from an external linguistics knowledge-base. The promising outcome as the result of Natural Language Processing (NLP) and Information Retrieval (IR) fields being merged together convinces us to create the extended ontology using NLP technique.

Keywords: Web ontology, WordNet, semantics relationships, automatic topic identification.

1 INTRODUCTION

The growing amount of information available in Internet has attracted many IR researchers to focus their works on web documents. In IR, most of the document categorization or classification use *bags-of-words*[1] [22] to represent the documents that needed to be categorized or classified. This kind of representation is not very suitable to be used to analyze the web document since a web document is more complex compared to the standard document (e.g Journals, Newspaper or Reports).

In our approach of creating the representation of a web document, the extracted keywords from the given web document is transformed into a web ontology related concepts. We believe that the content of the web document is best represented by these related concepts because with these related concepts, the semantic relations found among the keywords can be captured.

For example, if the keywords of the document are *computer* and *security*, one of the mapped path found in the ontology could be *Yahoo: Computer and Internet: Security and Encryption* (see figure1).

The transformation or mapping process will retrieve *Computer and Internet* and *Security and Encryption* concepts. The ontology helps us to identify that the word *security* mentioned in the web document is most probably talking about *computer security* which is related to *hackers* rather than *computer robbery*. However, a limited vocabulary of the web ontology will not be able to represent all or most of the document keywords. Therefore, we try to incoporate an external linguistics knowledge-base (WordNet) to enrich the ontology concepts. The words obtained from WordNet which are used to enrich the web ontology concepts are the extended ontology. This is the solution for the limited vocabulary problem and the same time provides us the alternative mapped concepts for the document keywords.

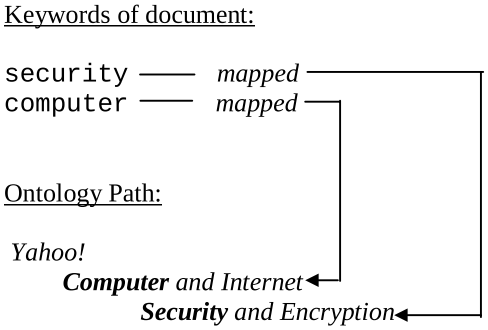

Figure 1: Transformation of keywords into web ontology concepts.

Therefore, our main goal in this paper is to find the best way on how the words from WordNet and the word concepts from the web ontology can be linked. With the best linkage established, we hope that the extended part of the ontology is true and will improve the performance of the topic identification module used in IR system.

We will briefly discuss about the related works on how to relate or to link two different sources

[1] A list of words extracted out from a document and used as the document representation.

of words in section 2. In section 3, the process of creating the extended ontology will be descibed. The experiment results will be in section 4 followed by our conclusion in section 5.

2 RELATED WORKS

In order to build our extended ontology, we will focus on the earlier related works that concentrate on linking a word with other words using a clearly defined relationship. Generally, there are many ways to link two or more different words. Some of them are using statistical words model, word probabilities model and lately many researchers have shown interest in NLP technique.

In this paper, we focus our method of linking two different sources of words by using NLP technique. But, before we describe our work further, we have to get a clear view on the impact of NLP technique towards this problem by looking at others' works.

In [27]'s paper, the effectiveness of the IR process is done by adding words that have lexical relationships to the query vector. Based on the experimental results, the most effective way of improving the IR accuracy performance is by adding synonym words and direct related synset words that have relative weight (α) of 0.5. [1] focused on the words role in order to make the user's query becomes flexible. By modifiying one or all the words in the user's query with synonym words, alternative queries can be produced to prevent a null retrieval result. However, in [18], they claimed that using other added words besides synonym words and root words on the user's original query words can mislead the retrieval process.

3 EXTENDED ONTOLOGY

We chose Yahoo as our web ontology based on the fact that Yahoo is the largest subject-directories of web documents and manually built with human knowledge towards Internet [26]. Our external linguistics knowledge-base is WordNet. WordNet is developed based on the theory of psycholinguistics by a group researchers from Princeton University [20]. In this linguistic knowledge-base, we can find words semantically related with the other words in many ways. We try to take advantage of these semantics relationships to establish links between the words of Yahoo concepts and words from WordNet. Based on the review of past related works [1],[28],[18], we decided to use three types of semantics relationships found in the WordNet and they are (with their definitions):

1. Synonym: Two words are synonymous if the substitution of one for the other never changes the truth value of a sentence in which the substitution is made [20]. In WordNet, synonymous words are put a group called synset.

2. Hypernym/hyponym: A "is-a" relation between words meanings [20]. Also known as a superset/subset relationship.

3. Meronym/holonym: A "part-whole" or "part-of" (or HAS-A) relationship between words [20].

Using these three semantics relationships, we can retrieve words from WordNet based on the words concepts of Yahoo. The retrieved words from WordNet will be treated as the extension of Yahoo ontology.

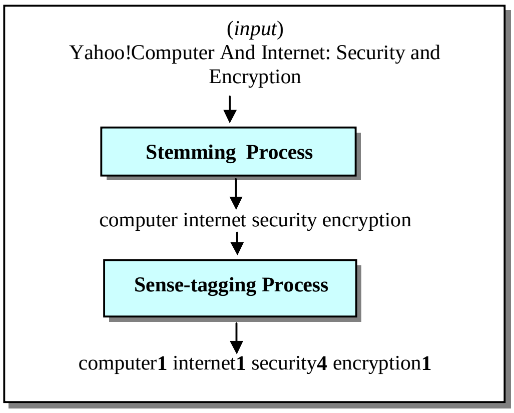
(*input*)
Yahoo!Computer And Internet: Security and Encryption
Stemming Process
computer internet security encryption
Sense-tagging Process
computer**1** internet**1** security**4** encryption**1**

Figure 2: Stemming process will change that words into root forms and sense-tagger process will give sense numbers to the stemmed words.

The enrichment of Yahoo concept using WordNet words is not as easy as seen. This is because the words exist from both resources are not the same. Words in WordNet are sense-tagged whereas the words of Yahoo concepts are ambiguous. Therefore we have to disambiguate these words of Yahoo concepts according to WordNet sense numbers using a sense-tagger system [23]. The sense-tagger system will take an ambiguous sentence as an input and produces an output of a stemmed and sense-tagged sentence. Figure 2 shows how the words concepts are sense-tagged.

After we have changed the form of the words in concept to be similar with the words in WordNet, we will try to get the words from WordNet using the three semantics relationships mentioned earlier. Figure 3 will give a clear view on how these three semantics relationship are used in order to obtain words from WordNet.

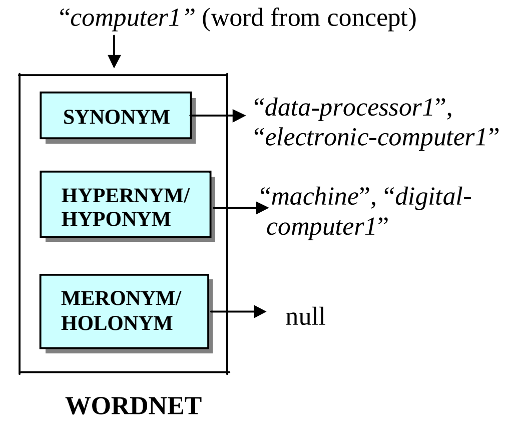
"computer1" (word from concept)
SYNONYM
"data-processor1", "electronic-computer1"
HYPERNYM/ HYPONYM
"machine", "digital-computer1"
MERONYM/ HOLONYM
null
WORDNET

Figure 3: Word *computer1* is used to obtain words from Wordnet using the three semantics relationships**.**

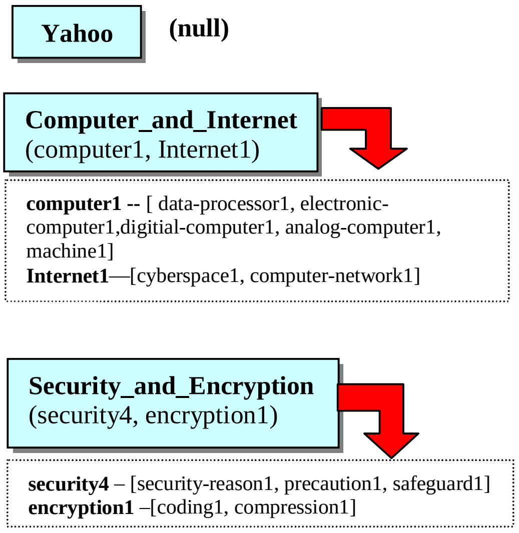
Yahoo
(null)
Computer_and_Internet
(computer1, Internet1)
computer1 -- [data-processor1, electronic-computer1,digitial-computer1, analog-computer1, machine1]
Internet1—[cyberspace1, computer-network1]
Security_and_Encryption
(security4, encryption1)
security4 – [security-reason1, precaution1, safeguard1]
encryption1 –[coding1, compression1]

Figure 4: The extended part (the box with broken line below the arrow) of each concept is built when semantically related words have been obtained from WordNet.

In figure 4, the words from WordNet will be linked to a concept where the word concept comes from. In the example, *computer1* comes from the concept of ***Computer_and_Internet***.

Whenever there is a failure to map the keywords of the document onto the ontology concepts, the system will look for alternative concepts from the extended ontology. If, in any case, the mapping process is not successful using both ways, then the keyword will be ignored and assumed that the keywords are not relevant to the content of the document.

4 RESULT AND ANALYSIS

For our experiment, we try to compare which extended ontologies that can be used to improve our automatic topic identification system [26]. The comparison is done by running the automatic topic identification system with two types of extended ontologies and also without extended ontology. The two types of ontologies are:

- Ontology based on synonym relationship.
- Ontology based on synonym, hypernym/hyponym and meronym/ holonym relationships.

Table 5 summarizes the results that we obtained from the experiment.

Extended Ontologies	meta-topic[2]	single path[3]	topic node[4]	total doc. analyzed
none	**68.5%**	**51.50%**	**29%**	**95.5%**
synonym	**69%**	**53%**	**30%**	**98%**
synonym+ hypernym/ hypnonym+ meronym/ holonym	**69%**	**52.65%**	**29%**	**100%**

Table 5: Results of two types of extended ontologies and one test without extended ontology.

We measure the performance of the automatic topic identification system in two different aspects:

- Precision: hits/ (hits + mistake).

[2] The parent node of the topic node at level two.

[3] The path where the topic node should be located.

[4] Topic node is the topic of the web document.

- Total document anaylzed =
  (total doc. hits +total doc.mistake)/
  (total all test documents)

The *precision* will measure how accurate the system identify the correct meta-topic, single path and topic node. The *total document* analyzed is the percentage of total documents that have been analyzed successfully. This means at least one keyword has been mapped onto the ontology concept.

With 200 test web documents and 107 nodes of Yahoo concepts, only 95.5% documents have been successfully represented by the ontology concepts. This percentage increased up to 98% when we included an extended ontology built based on synonym relationship. 100% is obtained when we enhanced the extended ontology with two more semantics relationships; hypernym/hyponym and meronym/holonym.

In terms of the *precision* measure, the result becomes worse when the extended ontology is built with more semantics relationships. As we can see, the improved result of 30% using synonym extended ontology declined to 29% when more semantics relationships are added.

4.1 ANALYSIS

Our accuracy result (precision on topic node) is quite comparable to the results produced by the other topic identification systems that use *bag-of-words* as the representative of the web documents. [19] only obtained 37% of accuracy using Yahoo with 151 classes and 50 documents. Their best attained result was 45% on 100 test documents. [9] had the worst result with only 2.13% at their preliminary experiment. Later, after they had increased the size of the sample training, the result improved to 36.5% accuracy (average).

The poor result we obtained with only 30% of maximum accuracy were caused by many factors. One of them could be because of the highly heterogenity of the web document. Therefore we were unable to extract the correct keywords from the web documents. The extraction of wrong keywords or less number of keywords is also due to the fact that some of the web documents are poorly written with spelling mistakes, non-standard language (slang) and mixture of other languages.

The situation of having worse result when the extended ontology was built based on more semantics relationships is not very surprising. This is because in [18]'s paper regarding query expansion using NLP technique, they claimed that other than using synonym words for query expansion, the true meaning of the original query will change and therefore can cause wrong information to be retrieved.

5 CONCLUSION

In this paper, we presented and evaluated our NLP approach in extending a web ontology. These web ontology and extended ontology will be used to build a document representative in automatic topic identification system. The process of merging the words from WordNet with Yahoo! ontology to build the extended part of Yahoo! ontology is done using NLP technique where semantics relationships defined in WordNet will be exploited.

Our main conclusion is that not all words which are semantically related to the words concepts of Yahoo are suitable to be used as the extended ontology. Some can even worsen the accuracy even though the number of keywords mapped onto the concepts has increased. Based on the result we have attained, building extended ontology based on synonym relationship is the best choice among all.

For future works, what we would like to do is find out whether, it is possible to achieve a high performance in both *Precision* and *Total document analyzed* by applying an appropriate weighting rule to the ontology concepts.

References

1. Banerjee, S., Mittal, V.O.: *On the Use of Linguistics Ontologies for Acessing Distributed Digital Libraries*. Proceeding of the First Annual Conference on Theory and Practice of Digital Libraries (1994)
2. Chakrabarti, S., Dom, B., Indyk, P.: *Enhanced Hypertext Categorization Using Hyperlinks*. ACM SIGMIND, Seattle, Washington (1998)
3. Chekuri, C., Goldwasser, M.H, Raghavan, P., Upfal, E.: *Web Search Using Automated Classification*. Poster at the Sixth International World Wide Web Conference (WWW6) (1997)
4. D' Alessio, D., Murray, K., Schiaffino, R., Kreshenbaum, A.: *Hierarchical Text Categorization*. Proceeding RIAO2000 (2000)
5. D' Alessio, D., Murray, K., Schiaffino, R., Kreshenbaum, A.: *The effect of Topological*

Structure on Hierarchical Text Categorization. Proceeding of the Sixth Workshop on Very Large Corpora, COLLING ACL '98 (1998)

6. Gelbukh, A., Sidorov, G., Guzman, A.: *A Method of Describing Document Contents through Topic Selection*. In Proc. of International Symposium on String Processing and Information Retrieval, Cancun, Mexico. Library of Congress 99-64139, IEEE Computer Society Press (1999)
7. Gelbukh, A., Sidorov, G., Guzman, A.: *Use of a Weighted Topic Hierarchy for Document Classification*. In Václav Matoušek et al (eds.): Text, Speech and Dialogue in Poc. 2nd International Workshop. Lecture Notes in Artificial Intelligence, No.92, ISBN 3-540-66494-7, Springer-Verlag., Czech Republic (1999) 130-135
8. Gelbukh, A., Sidorov, G., Guzman, A.,: *Text Categorization Using a Hierarchical Topic Dictionary*. Proc. Text Mining Workshop at 16th International Joint Conference on Artificial Intelligence (IJCAI'99), Stockholm, Sweden (1999)
9. Gövert, N., Lalmas, M., Fuhr, N.: *A Probabilistic Description-Oriented Approach for Categorizing Web Document.* Proceeding of the Eighth International Conference on Information Knowledge Management, Kansas City, MO USA (1999) 475-482
10. Greiner, R., Grove, A, Schuurmans, D.: *On learning hierarchical Classifications* . In ResearchIndex; The NECI Scientifc Literature Digital Libraray [Online]. Available from: http:// citerseer. nj.nec /com / 38202. html [Accessed 25 July] . (1997)
11. Grobelnik, M., Mladenic, D.: *Fast Categorization*. In Proceedings of Third International Conference on Knowledge Discovery Data Mining (1998)
12. Guzman, A.: *Finding the Main Themes in a Spanish Document*. Journal Expert Systems with Application (1998) 139-148
13. *Hoenkamp, E.: Spotting Ontological Lacunae through Spectrum Analysis Of Retrieved Documents*. 13th European Conference On Artificial Intelligent, ECAI98, Brighton, England (1998)
14. Koller, D., Sahami, M.: *Hierarchically Classifying Documents Using Very Few Words*. In the Proceeding of Machine Learning (ICML-97) (1997) 170-176
15. Lee, J. Shin, D.: *Multilevel Automatic Categorization for Webpages*. The INET Proceeding '98 (1998)
16. Lin, C.Y, Hovy, E.: *Identifying Topics by Position*. In the Proceeding of The Workshop of Intelligent Scalable Text Summarization '97 (1997)
17. Lin, C.Y: Knowledge-based Automatic Topic Identification. In the Proceeding of The 33rd Annual Meeting of the Association for Computational Linguistics '95 (1995)
18. Loupy, C. de., Bellot, P., El_Beze, M. & Marteau, P. F.:*Query Expansion and Classification of Retreived Documents*. In the Proc. of the 17th Text REtrieval Conference, NIST Publication, Gaithersburg, Maryland (1998)
19. McCallum, A., Rosenfeld, R., Mitchell, T., Ng, Y.A.: *Improving Text Classification by Shrinkage in a Hierarchy of Classes*. Proceeding of the 15th Conference on Machine Learning (ICML-98) (1998)
20. Miller, G.A, Beckwith, R., Fellbaum, C., Gross, D., Miller, K.: *Introduction to WordNet: An-Online Lexical Database*. Five Papers on WordNet (1993)
21. Quek, C.Y, Mitchell, T: *Classification of World Wide Web Documents*. Seniors Honors Thesis, School of Computer Science, Carnegie Melon University (1998)
22. Scott, S., Matwin, S.: *Text Classification using WordNet Hypernyms*. In the Proceeding of Workshop – Usage of WordNet in Natural Language Processing Systems, Montreal, Canada (1998)
23. *Sense Tagger*. UTMK Internal Paper. Universiti Sains Malaysia, Penang, Malaysia (1999)
24. Soderland, S.: *Learning to extract text-based information from World Wide Web*. In the Proceeding of the Third International Conference on Knowledge Discovery and Data-Mining (1997)
25. Solock, J.: *Searching the Internet Part II: Subject Catalogs, Annotated Directories, and Subject Guides* [Online]. Available from: http://rs.internic.net/nic-support/nicnews/oct96/enduser.html [Accessed 3 March 1999]
26. Tiun, S., Abdullah, R. & Tang, E.: *Automatic Topic Identification Using Ontology Hierarchy*. In the Proc. of the 2nd International Conference on Intelligent Text Processing and Computational Linguistics (CICLing-2001), Mexico City, Mexico (2001).
27. Voorhees, E.M.: *On Expanding Query Vectors with Lexically Related Words*. Proceeding of the Second Text REtrieval Conference (TREC-2), NIST Special Publication, Gatherburg, Maryland (1993)