Automatic Topic Identification Using Ontology Hierarchy

Sabrina Tiun[1], Rosni Abdullah[2], Tang Enya Kong[3]

[1]UTMK, P.Pengajian Sains Komputer, Universiti Sains Malaysia, 11800
Pulau Pinang, Malaysia
sab@cs.usm.my

[2]Pusat Pengajian Sains Komputer, Universiti Sains Malaysia, 11800 Pulau
Pinang, Malaysia
rosni@cs.usm.my

[3]UTMK, P.Pengajian Sains Komputer, Universiti Sains Malaysia, 11800
Pulau Pinang, Malaysia
enyakong@cs.usm.my

Abstract. This paper proposes a method of using ontology hierarchy in automatic topic identification. The fundamental idea behind this work is to exploit an ontology hierarchical structure in order to find a topic of a text. The keywords which are extracted from a given text will be mapped onto their corresponding concepts in the ontology. By optimizing the corresponding concepts, we will pick a single node among the concepts nodes which we believe is the topic of the target text. However, a limited vocabulary problem is encountered while mapping the keywords onto their corresponding concepts. This situation forces us to extend the ontology by enriching each of its concepts with new concepts using the external linguistics knowledge-base (WordNet). Our intuition of a high number keywords mapped onto the ontology concepts is that our topic identification technique can perform at its best.

1 Introduction

The growing amount of information available in Internet has attracted many researchers to focus their works on text analysis and processing on web document. One of the important process is to find the main topic for a particular web document. Our proposed approach of identifying the main topic for a web document is by exploiting a web ontology hierarchical structure. In order to do this, we have to find a way to map the text representation onto the ontology concepts. We intuitively believe that the content of the text is better represented by these related mapped concepts. Using these related concepts, we can capture the semantics relation found among the words in the text. For example, if the extracted words from a web document are *computer* and *security*. Below is one of the mapped paths found in the ontology hierarchy:

Yahoo!Computer and Internet:Security and Encyption

The mapping process will retrieve both *Computer* and *Security and Encryption.* The ontology hierarchy helps us to identify that the security mentioned in the web document is most probably talking about *computer security* which is related to *hackers* rather than *computer robbery*. However, the limited vocabulary of the web ontology is unable to represent all or most of the document keywords. Therefore, we try to incorporate an external linguistics knowledge-base (WordNet) to enrich the ontology concepts. This is what we call as the extended ontology hierarchy because each of the node concepts in the ontology will be extended with words obtained from WordNet. This is the solution for the limited vocabulary problem and at the same time provides us the alternative mapped concepts for the text keywords.

We will briefly discuss the related works regarding the topic identification using hierarchical structure in section 2. The process of creating the extended ontology will be in section 3. In section 4, we will describe the whole process of our automatic topic identification system. The experiment results will be in section 5 followed by our conclusion in section 6.

2 Related Works

We found that the topic identification based on hierarchical structure technique is usually applied in text classification system [21], [14], [3], [11], [4], [24], [15], [18], [7]. Most of these works utilize a hierarchical structure to decompose a huge and single classifier task into a set of small classifiers that correspond to the node categories of the hierarchy. By placing a classifier at each node, a set of features is extracted at each node category. The topic of the new document will be identified by computing the similarity between the document feature and each of the node categories feature or the probability that the document belongs to a node category [14]. Other methods are like [21] that emphasise on representing a document with hypernym density using the WordNet hypernym hierarchy and [15] which combines clustering method with text categorization.

Our approach of identifying the topic of a document has been inspired by the works of [21] and [17]. [17] extends the word frequency counting (the classic way of identifying text topic) to concept counting. He uses the WordNet taxonomy to collect interesting concepts and later generalizes the concepts to identify the main topic of the target text. In [21], they transform the bag-of-words representation of the text into hypernym density. Using the height of generalization, they can limit the number of steps upward through the hypernym hierarchy for each word.

The most comparable works to our technique of topic identification are by [6], [7] and [12] who uses a concept tree which was built manually (in our work, we use existing Yahoo ontology). In [12], the program called CLASITEX will analyze a document and look for words or collocations appearing to be the tree concepts. The concepts will be counted over the concept tree and this counting includes the upper level concepts. At the end, some parts of the concept tree will have higher counts and those parts are most likely the topics of the given document. In [6] and [7], they use numerical weight in measuring the relevance of the words for topics and the important of the nodes of the hierarchy to determine the principle topics of a given document. This method is implemented in a system called CLASSIFIER**.**

3 Extended Ontology

We choose 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 toward Internet. Our external linguistics database is WordNet. WordNet is developed based on the theory of psycholinguistics by a group of researchers from Princeton University [19]. 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 WordNet vocabulary. Based on our review on past related works [1], [13], [24] we decide to use three types of semantics relationships found in the WordNet and they are synonym[1], hypernym/hyponym[2] and meronym/holonym[3]. Using these three semantics relationships, we can retrieve words from WordNet based on the words of Yahoo concepts. These retrieved words from WordNet will be treated as the extension of Yahoo hierarchy.

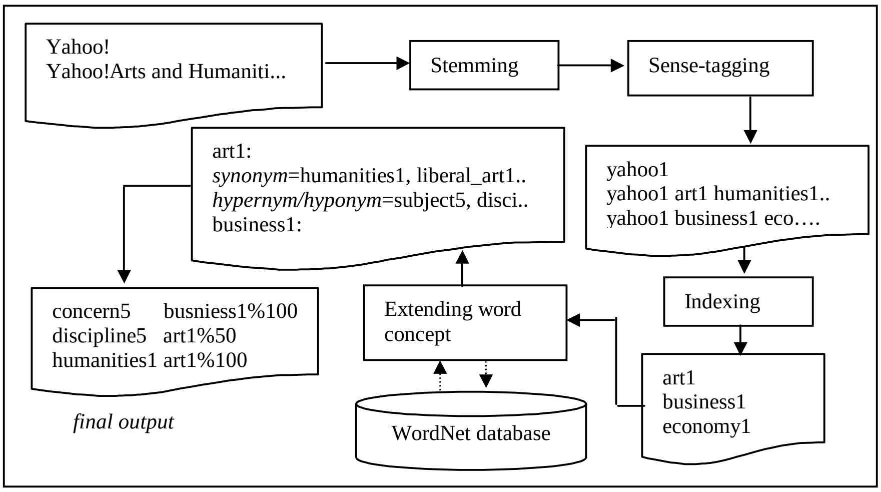

Fig. 1. The development process of the extended ontology flowchart. In the final output file, the words without the percent symbol are the extended concepts

The enrichment of Yahoo concept using WordNet words is not as easy as seen. This is because the words which exist in both sources 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 concept according to WordNet sense numbers using a sense tagging process [22]. Basically, what we have to do for the sense-tagging process is to prepare a “dictionary file”. This dictionary file contains all words found in the ontology domain with all the possible word

[1] Two words are equivalent.
[2] This relationship is also known as “is-a” or “superset/subset’ relationship.
[3] This relationship is also known as “has-a’ or “part-of” relationship.

senses found in the WordNet (together with the word senses synonyms, intermediate parents and definitions). Later, the dictionary will be used in constructing a metric of senses relatedness file for the sense-tagger process. This metrics file is used to calculate how close words are related in the input sentence.This sense-tagger process will take only one ambiguous sentence as an input at one time and produce an output of a stemmed and sense-tagged sentence.

When the word concepts have been sense-tagged, then they are ready to be enriched with WordNet words. Fig.1 shows the process on how we build the extended part of the ontology.

4 Automatic Topic Identification

Generally, our automatic topic identification system has three main components: The extraction module, mapping module and optimization module. The input of the system is a web document and the output will be a node concept which is also the predicted topic of the target document. The node concept can be in a form of one word or more. Fig.2 shows the general overview of our topic identification system.

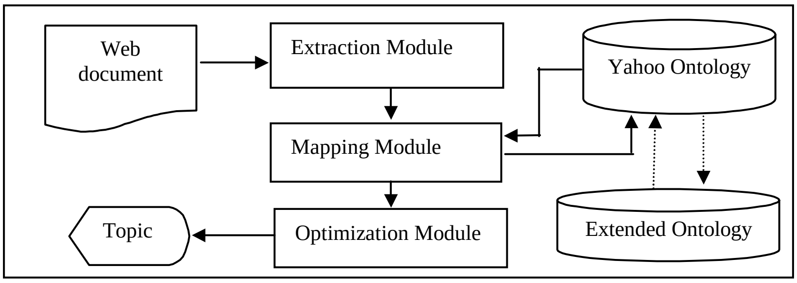

Fig. 2. The general overview of the automatic topic identification

4.1 Extraction Module

The extraction module handles the process of extracting important sentences from the document. Our method of extraction is based on the HTML tag. This is because we believe some of the HTML tags indicate the location where the authors may emphasise their ideas. For example, author may choose the best words to describe his web page at the title tag. Therefore we will choose sentences or words that become the pointers to other documents, words which are highlighted and words located in the title tag. However, some web documents maybe lack of HTML tags and therefore our extraction technique is not appropriate to be used. In this case, alternative way of extracting web document like words frequency [18], [3] ,[11] and positional policy [16] are more applicable. Since in this paper we are interested in extracting out information from the web document based on HTML tag, we consider extracting information from non-structured document as a future work.

When all the sentences have been extracted from the web document, these sentences will be stemmed and sense-tagged using the same sense-tagger used in the system that develops the extended part of the web ontology. This will ensure the consistency of word sense number used in our system. The final output of this module will be a list of keywords (from the extracted sentences) which are already stemmed and sense-tagged.

4.2 Mapping Module

The mapping module will take the output of the extraction module as an input. The keywords will be mapped on the words of ontology concepts that have been stemmed and sense-tagged. However, there is a possibility that the keyword may not be able to be mapped onto its corresponding concept because there is no such concept available in the Yahoo ontology. This situation requires an alternative way to map the keyword onto the Yahoo concept. The alternative way is to use the extended concept as a "middle man" in order the mapping between Yahoo concept and the keyword becomes possible.

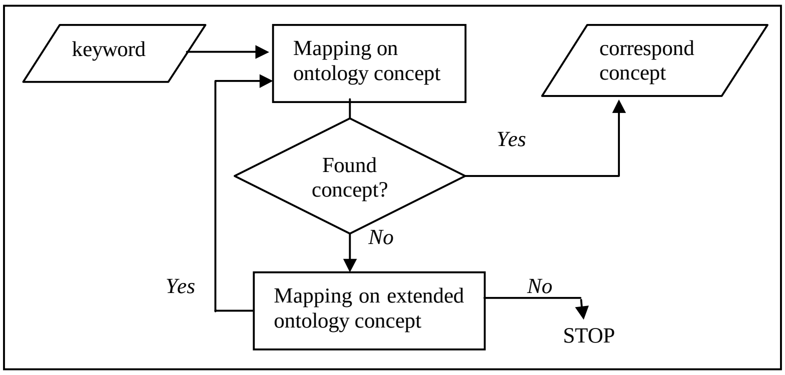
keyword
Mapping on ontology concept
correspond concept
Yes
Found concept?
No
Yes
Mapping on extended ontology concept
No
STOP

Fig. 3. The flowchart of the mapping and the alternative mapping between the keyword and the Yahoo ontology concept

4.2.1.1 The concept weight

In order to discriminate between the important concept and the less important concept, we need to give weight on the concepts. The determinants that influence a weight given to a node concept are the type of mapping and the number of keyword frequency. The keyword frequency indicates how frequent the particular concept is mentioned in the document. The higher the frequency, the more important the concept is deemed to be. However, given weight based on frequency should be re-checked because the type of mapping determines how accurate the keyword and the concept is mapped. Mapping using extended concept reduced the frequency of the concept because this kind of mapping is done indirectly. This kind of mapping weakens the confidence that the concept is the true representative of the keyword.

There are two types of weight that the node concept will have before we proceed on the process of identifying the node topic. The node concept will have its weight alone and its accumulated mixture weight. The sole weight of the node is calculated based on the word frequency and the way the document keyword mapped onto the corresponding node. The word frequency will be the number of the keyword appeared in the document.

When the node has this weight, the weight will be reduced if the type of mapping between the keyword and the node concept is indirect (the alternative mapping). The amount of weight being reduced will be depended on the semantics relationship that exists between the extended concept and Yahoo concept. In our implementation, all the mappings through the extended concept will reduce the node weight up to 50% except the extended concepts that are built based on synonym relationship.

For example, we have a concept *Arts and Humanities* that has *art1* and *humanities1* as the words that have been stemmed and sense-tagged. If the keyword of a document is *art1*, the keyword will have a direct mapping because *art1* is found in the node concept. Whereas, if the keyword is *subject5*, the keyword will be mapped indirectly to the *Arts and Humanities* concept. This is because *subject5* is found in the extended concepts of the *Arts and Humanities* (using hypernym/hyponym relationship). If the number of both keywords *art1* and *subject5* appeared in the document is five times, the weight contribution of word *art1* to the *Arts and Humanities* concept will be 7.5 (refer to the calculation below).

art1 {keyword}➔ art1 {*yahoo concept* } = 5 unit
subject5{*keyword*} ➔ subject5{*extended concept*} ➔ art1{ *yahoo concept* }
= 5 x 0.5 = 2.5

Here, we would like to remind that extended concepts will never be considered to be the candidates of the node topic but rather as the medium where keywords can be mapped onto the Yahoo concept. This is because we assume that the possible topic of a web document can only be found within the Yahoo ontology terminology.

In order to calculate the accumulated weight of a node, the weight of the children node will propagate upward. The accumulated mixture weight of a node will be the total of accumulated weight of the children node (including the node itself) times with the number of different words concept accumulated from the children node (including the node itself).

For an example, *Arts and Humanities:History*

art1 {*Arts and Humanities*} = 7.5 unit weight
history1 {*History*} = 2 unit weight

The accumulated weight of *Arts and Humanities* will be: 7.5 + 2 = 9.5, where 2 comes from *history,* the child node of *Arts and Humanities.* The accumulated mixture weight of concept *Arts and Humanities* will be: (7.5 + 2) x 2 = 19 unit. The 2 here comes from the total number of different words concepts which are *art1* and *history1*.

4.3 Optimization Module and Main Topic

The optimization process will shrink the ontology tree into an optimized tree where only active concepts and the intermediate active concepts are chosen. The small size of optimized tree will be reduced to a single path. This single path is retrieved using the Maximal Spanning Tree Algorithm (same as Minimal Spanning Tree Algorithm). The Maximal Spanning Tree algorithm will find a path that has the heaviest nodes. The weight of the node that the algorithm uses as the criteria to choose a heaviest node will be the accumulated mixture weight that the node has.

In order to find a node which is highly suspicious to be the node topic, we build an algorithm which can find the node that has the largest accumulated mixture weight distribution over the optimized tree. The algorithm is called *Ratio Balance Algorithm*. This algorithm will find the maximum ratio balance of the single path nodes by subtracting the *actual accumulated mixture weight* with the *supposed accumulated mixture weight*. The supposed accumulated mixture weight of a node is calculated by dividing the root node accumulated mixture weight with the level where the node is located. The level where root node located is equal to 1. For example, we have a single path consists of these nodes concepts, *Yahoo:Arts and Humanities:Histor*y and we have the calculation of ratio balance for each of the nodes in the next page:

Table 1. An example of Ratio Balance calculation

Level	Node Concept	Supposed acc. mix.weight	Actual acc.mix.weight	Ratio Balance
1	*Yahoo*	19	19	-
2	*Arts and Humanities*	19/2 = 9.5	19	19– 9.5 = 9.5
3	*History*	19/3 = 6.33	2	2 – 6.33 = -4.33

Since the node that has the maximum ratio balance will be the topic node, on the above example *Arts and Humanities* is the topic concept for the target document. Root node will be excluded because it is too general and useless (this is referred to *Yahoo* as the root node).

5 Experiment: Result and Analysis

We tested our system on 107 node categories with 202 web documents downloaded from *Yahoo!Business and Economy:Transportation* directory. We measured the system accuracy using a precision formula as below:

Precision = hits / (hits + mistakes)

The table below shows the result of accuracy on the meta-topic[4], single path and on topic node:

	Meta-topic	Single Path	Topic Node
Precision	69.8%	51.9%	29.7%

The system had extracted 1163 keywords from the 202 web documents and only 57.8% keywords were successfully mapped onto their correspond concepts through direct mapping. 17.8% of the keywords were mapped onto their correspond concepts through extended concepts. The system was unable to map 24.68% of the 1163 keywords onto the keywords corresponding concepts.

Analysis. Our result is quite comparable to that result produced by the topic identification (text classification) that used machine learning techniques on web documents classification.

[18] obtained 37% of accuracy result using Yahoo with 151 classes and 50 documents. Their best attained accuracy result was 45% on 100 test documents. [11] had their best result on accuracy at 51% using Yahoo hierarchy and Yahoo web documents. The non-hierarchical web documents automatic categorization by [9] that used probabilistic description-oriented had worst result on their preliminary experiment with only an average precision of 2.13%. Later after they had increased the size of their sample learning, their result on categorization accuracy increased up to 36.5%.

The poor result we obtained with only 29.7% of accuracy could be caused by many factors. One of them could be the highly heterogeneity of web document that we could not extract all the web document information. This includes the fact that some of the web documents are poorly written with spelling errors, non-standard language (slang) and written in many languages. The web document is also not well represented by our concepts because the mapping between the keywords and the ontology concepts is not 100% successful. Further more, this poor result is also contributed to the kind of model we use which is the top-down model. This model will face unrecoverable mistakes because if we already choose the wrong node at the top, our topic node will be wrong too. This problem is faced by most of the automatic text classification using hierarchical structure [14], [25], [10].

In this experiment we only use a small size of web ontology. This is because we are only interested in how our topic identification system performs on real data regardless of size. However, for future work, we intend to build our automatic topic identification prototype system in a distributed environment so that a large web ontology can be implemented. The distributed topic identification system will be designed to identify more than one topics per document.

[4] The node located one level below the root node.

6 Conclusion

In this paper, we presented and evaluated an approach that automatically identifies a topic of a web document by exploiting a web ontology hierarchical structure (Yahoo). In spite of that, we also showed how we enriched the web ontology concept using external linguistics database, WordNet. The merging of WordNet words with Yahoo concept was made by using three types of semantics relationships found in WordNet.

Our main conclusion is that our approach is simple yet quite comparable to the other complex methods (probabilistic and machine learning method). We need more time to work with our dictionaries for the sense-tagger system which resulted the unsuccessful mapping between the keywords and the concepts. We may also need to enrich the ontology concepts with more knowledge besides semantically related words in order to understand more about the content of the web documents. In future, we hope we can come out with a more efficient and improved version of this automatic topic identification system and implement it in other automatic text processing systems like text classification or in information retrieval systems.

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 (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. 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)
19. Miller, G.A, Beckwith, R., Fellbaum, C., Gross, D., Miller, K.: Introduction to WordNet: An-Online Lexical Database. Five Papers on WordNet (1993)
20. Quek, C.Y, Mitchell, T: Classification of World Wide Web Documents. Seniors Honors Thesis, School of Computer Science, Carnegie Melon University (1998)
21. 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)
22. Sense Tagger. UTMK Internal Paper. Universiti Sains Malaysia, Penang, Malaysia (1999)
23. 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)
24. 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)
25. Weigned, A.S, Wiener, E.D, Pedersen, J.O.: Working Papers IS-98-22. Dept. of Info. System, Leonard N. Stern, School Of Business, New York University (1998)