An Automatic Approach to Classify Web Documents Using a Domain Ontology

Mu-Hee Song, Soo-Yeon Lim, Seong-Bae Park, Dong-Jin Kang, and Sang-Jo Lee

Dept. of Computer Engineering, Information Technology Services,
Kyungpook National University, Daegu, The Korea
{mhsong, seongbae, djkang, sjlee}@knu.ac.kr
nadalsy@hotmail.com

Abstract. This paper suggests an automated method for document classification using an ontology, which expresses terminology information and vocabulary contained in Web documents by way of a hierarchical structure. Ontology-based document classification involves determining document features that represent the Web documents most accurately, and classifying them into the most appropriate categories after analyzing their contents by using at least two pre-defined categories per given document features. In this paper, Web documents are classified in real time not with experimental data or a learning process, but by similar calculations between the terminology information extracted from Web texts and ontology categories. This results in a more accurate document classification since the meanings and relationships unique to each document are determined.

Keywords: Document classification, Ontology, Web Page classification.

1 Introduction

In recent years ontologies have become a topic of interest in computer science. This paper suggests an automated method for document classification using ontology. In our research, Web documents are classified based on similarities determined by the ontology, which expresses the meaning structure of the Web documents' terminology information and vocabulary in a hierarchical manner. Identifying and comparing the meaning content and relationship of each document can perform document classification more accurately and efficiently. The ontology mentioned in this paper is comprised of concepts, concepts features, relations between concepts, and constraints for document classification, all in a hierarchical manner. Also, the ontology's hierarchical structure is applied to the document classification.

Our work is distinguished from others for the following reasons: (1) Rather than using a dictionary or knowledge index, ontology is used for document classification. (2) Our ontology is based on syntax information contained in the Web pages. (3) Mapping between the established ontology and terminology information extracted from Web pages is performed.

We wish our classification system could be used to classify current web pages into proper categories and to generate reports if the web pages belong to unwanted sites.

S.K. Pal et al. (Eds.): PReMI 2005, LNCS 3776, pp. 666 – 671, 2005.

In the following section of this paper, we will discuss related approaches to our classifier. Section 3 describes the details of our framework how ontology, as suggested in this paper, is applied to document classification, and we show the experiments and evaluation of web page in Section 4. In conclusion we summarize our results and mention further research issues..

2 Related Works

This section describes research which has been carried out by people in the area of automatic document classification, and it examines the key difference between our work and other research.

The rule-based model [1] utilizes experts' help based on generally distinguished rules that appear in study texts or applies rules that are extracted by studying the documents. The Bayesian probability model [2] applies the probability theory to the document features extracted from the documents. The SVM (Support Vector Machine: SVM)[3] uses the machine learning method. Although these methods have some degree of accuracy, all of them require some rule learning level and they must have the training data as a reference.

There have been several approaches which focus on ontologies to classify Web pages [4,5,6]. Prabowo et. al. [4] defined ontology as "a single entity which holds conceptual instances of a domain, and differentiates itself from another," and used ontologies for web page classification with respect to the Dewey Decimal Classification(DDC) and Library of Congress Classification(LCC). The weakness of their approach is the fact that is not adaptive when users require more sophisticated classification, even if the approach follows the standard classifications. However, since our approach builds an adaptive ontology, we provide a flexible classification reflecting requests.

This paper suggests an ontology-based, automated document classification method, which does not require these learning processes and can be performed in real time.

3 Document Classification Using Ontology

3.1 Ontology Structure

In this paper, ontology is defined as one independent, collective representation of all standardized concepts for vocabulary and terms in one place. Here, we are not talking about collections of simple words, but we are referring to collection of vocabulary, which have relationships with both simple rules and meanings. Ontology expressions are based not only on the logical relations between term definitions and other meanings, but also on the bottom-out structure where the interpretation starts from primitive terms. We have decided to apply ontology to Web document classification, because it has the unique, hierarchical structure and characteristic of machine reasoning, starting from very primitive terms.

The advantages of an ontology-based classification approach over the existing ones, such as hierarchical[7], - and probabilistic – approach[8], are that (1) the nature of the relational structure of an ontology provides a mechanism to enable machine reasoning; (2) the conceptual instances within an ontology are not only a bag of keywords but have inherent semantics, and a close relationship with the class representatives of the classification schemes. Hence, they can be mapped to each other; (3) this is a kind of Web page and class representative. It also enables us to get insights into and observe the way the classifier assigns a class representative to Web pages by tracking the links between the conceptual instances involved and the associated class representative [4].

3.2 Building Domain Ontology for Document Classification

In this research, ontology for the ‘economy’ domain has been developed for experimental purposes of document classification. To configure and develop the ontology, it is first assumed that vocabulary which frequently appears in document collection is similarly related to other vocabulary. The second point is this frequently appearing vocabulary is used to build the basic network structure. Third, adding vocabulary that has a relationship with those selected words expands the ontology. Then, similarities between the terminology information extracted from Web page, and ontology terminology data are identified and compared in order to start the document classification process.

3.3 Document Classification Using Ontology

The process of Web document classification basically involves two procedures: Finding key vocabulary in the documents and mapping onto a node in the concept hierarchy (ontology) using the extracted words.

Table 1. The document-term frequency data matrix after the stemming and stopping processes

Doc_j	**TF_1**	**TF_2**	**…**	**TF_m**
Doc_1	2	4	…	5
Doc_2	2	3	…	2
Doc_3	2	3	…	2
…	…	...	...	...
Doc_n	1	3	…	7

As part of the key vocabulary extraction process from documents, the removal of stop words and the stemming of words, both as the pre-classification procedures as well as the application of information retrieval measurement, *tf*×*idf* (term frequency times inverted document frequency), take place. After the stemming and stopping process of the terms in each document, we will represent them as the document-term frequency matrix ($Doc_j \times TF_{jk}$) as shown in Table 1. Doc_j is referring to each web page document that exists in the news database where $j=1,....,n$. Term frequency TF_{jk} is the number of how many times the distinct word w_k occurs in document Doc_j where $k=1,...,m$. The calculation of the terms weight x_{jk} of each word w_k is done by using a method that has been used by Salton[9][10] which is given by

$$x_{jk} = TF_{jk} \times idf_k \tag{1}$$

The similar calculation for classification is done using the following formula: The text is mapped onto a node with the highest similarity value, and one text is ultimately classified into one class.

$$Sim(Node, d) = \frac{\sum_{i=0}^{N} freq_{i,d} / \max_{i,d}}{N} \times \frac{V_d}{V} \qquad (2)$$

N is the feature frequency of a node. $freq_{i,d}$ represents the frequency of feature *j* that is matched in text *d*. $max_{i,d}$ is the frequency of the feature that is matched the most in Text *d*. *V* is the number of constraints, while V_d represents the number of constraints that are satisfied by Text *d*. The document classification takes place only when the use of the relations is "is-a", "has-a", "part-of", or "has-part". When another node is related, it is also included in the classification process to calculate the similarity. Using this approach, a more accurate classification can be performed.

The overall classification process for Web documents is as shown below (Figure 2).

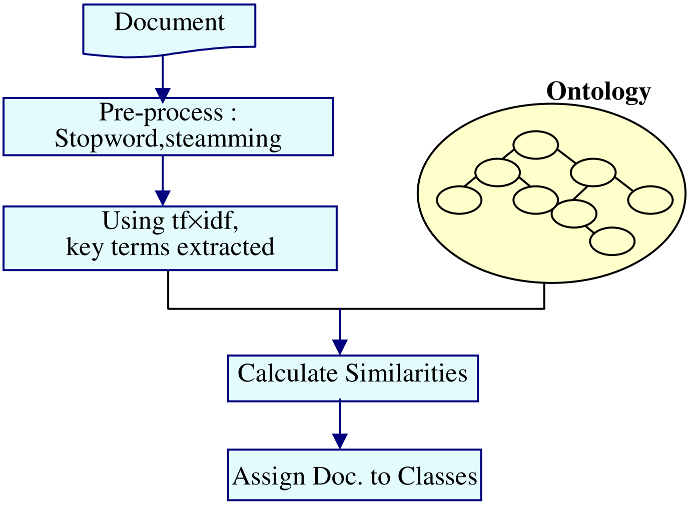

Fig. 2. Web Document Classification Process Using Ontology

4 Experimental Procedures

We have used a web pages dataset from Yahoo Economy news as shown in Table 2. The types of news in the database are Cooperatives, Employment, Finance, Marketing, Organizations and Trade. The total of documents are 5,235.

Automatic classification of web page is evaluated using the standard information retrieval measures that are precision, recall, and *F1* [11]. The *F1* measure is a kind of average of precision and recall.

Table 2. The number of document that are stored in the news database

Class no.	Class name	Number of Doc
1	Cooperatives	620
2	Employment	1,685
3	Finance	750
4	Marketing	680
5	Organizations	650
6	Trade	850
	Total	5,235

Table 3. The classification results

Class no.	Precision (%)	Recall (%)	F1 (%)
1	77.21	93.84	84.72
2	92.48	94.16	93.31
3	93.93	95.38	94.65
4	91.17	95.38	93.23
5	91.30	96.92	94.03
6	91.97	96.92	94.38
Average	**89.68**	**95.43**	**92.39**

(*These are the classes that exist in Economy category)

In Our the classification results, the precision, recall, and *F1* measures are 89.68%, 95.43%, 92.39%, respectively, as shown in Table 3.

A better document selection approach needs to be used for selecting the candidate documents from each class in order to increase the *F1* classification results.

5 Conclusion and Future Research

This paper introduced the use of ontology to conceptually express the meaning of relationships contained in Web documents and suggested an automated document classification method using the ontology. In particular, this paper focused on document classification based on the similarities of documents already categorized by ontology using terminology information extracted from the documents. Our work is distinguished from other studies in the following areas. (1) Rather than using a dictionary or knowledge index, ontology is used for document classification. (2) Our ontology is based on syntax information contained in the Web texts. (3) Mapping between the established ontology and the term information extracted from Web documents is performed. The document classification technique proposed by this paper does not involve any learning processes or experimental data and can be performed in real time.

Further research is required to develop more efficient and accurate ontological expressions and to document classification methods. We plan to conduct further studies on how to improve the efficiency of an information search using the document classification technique suggested in this paper and how to automatically determine the meaning of concepts and relations from Web documents.

References

1. Chidanand Apt, Fred Damerau, and Sholom M. Weis, "Towards Language Independent Automated Learning of Text Categorization models," *Proc. of the 17th annual international ACM-SIGIR, 1994.*
2. R.E.Shapire, Yoram Singhal, and Amit Singhal, "Boosting and Rocchio applied to text filtering,", *Proc. Of the 21th annual international ACM-SIGIR, 1998.*
3. Mart A. Hearst, "Support Vector Machines," *IEEE Information Systems, 13(4):18~28, 1998.*
4. Rudy Prabowo, Mike Jackson, Peter Burden, and Heinz-Dieter Knoell, 2002, "Ontology-Based Automatic Classification for the Web Pages:Design,Implementation and Evaluation,", *Proc. Of the 3rd International Conference on Web Information Systems Engineering, 2002.*
5. C.Jenkins, M.Jackson, P.Burden, and J.Wallis, "Automatic RDF metadata generation for resource discovery", *Proc. Of 8th International WWW Conference, Toronto, pp. 11-14, May 1999.*
6. Y.Ng, J.Tang, and M.Goodrich, "A binary categorization approach for classifying multiple-record Web documents using application ontologies and a probabilistic model", *Proc. Of 7th International Conference on Database Systems for Advances Applications, pp.58-65, April 2001.*
7. S.T.DUMAIS , and H.CHEN,"Hierarchical classification of Web content,", *Proc of the 23rd Annual International ACM SIGIR, July 24-28,2000, Arthens, Greece.*
8. N.GOEVERT, M.LALMAS, and N.FUHR, 1999, "A probabilistic description-oriented approach for categorisiong Web documents,", *Proc. Of the 8th ACM International Conference on Information and Knowledge Management, November 2-4, 1999,pp 475-482, Kansas City, U.S.*
9. Salton&McGill, Introduction to modern information retrival, New York, Mcgraw-Hill, USA, 1983.
10. Andreas Hotho and Alexander Maedche and Steffen Staab, "Ontology-based Text Document Clustering", Http://www.aifb.uni-karlsruhe.de/WBS
11. D.D.Lewis, Evaluating and optimizing autonomous text classification systems, in: E.A.Fox, P.Ingwersen, R.Fidel(Ed.), *SIGIR'95: Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrival*, New York, 1995, pp. 246-254.