International Journal of Computational Intelligence and Applications

IDENTIFYING TOPICS FOR WEB DOCUMENTS THROUGH FUZZY ASSOCIATION LEARNING

CHOOCHART HARUECHAIYASAK, MEI-LING SHYU

Department of Electrical and Computer Engineering,
University of Miami, Coral Gables, FL 33124-0640, USA

SHU-CHING CHEN

Distributed Multimedia Information System Laboratory, School of Computer Science,
Florida International University, Miami, FL 33199, USA

Received (November 30, 2001)
Revised (May 31, 2002)

Due to the explosive growth of available information on the World Wide Web (WWW), users have suffered from the information overload. To alleviate this problem, there is a need for an intelligent tool to help the users screening and filtering for interesting and useful information. In this paper, a method of automatically identifying topics for Web documents via a classification technique is proposed. Topic identification can be applied as a filtering tool for recommender systems to prune down the number of documents to within some particular topics. We adopt the fuzzy association concept as a machine learning technique to classify the documents into some predefined categories or topics. Our approach is compared to the vector space model with the cosine coefficient using the data sets collected from three different Web portals: Yahoo!, Open Directory Project and Excite. The results show that our approach yields higher classification accuracy compared to the vector space model.

Keywords: Topic identification; text mining; information filtering; document classification; fuzzy association learning.

1. Introduction

With the amount of information growing at an exponential rate, the World Wide Web (WWW) is often referred to as the world's largest and fastest growing information source[1]. It is not uncommon that the users on WWW often find themselves overwhelmed with the large amount of information that might be of their interest and usefulness. This problem is generally known as the information overload. To alleviate the problem, many data mining techniques have been applied into the Web context. This research area is generally known as Web mining[2]. Web mining is defined as the discovery and analysis of useful information from WWW. Some examples of Web mining techniques include analysis of user access patterns[3,4], Web document clustering[5,6], classification[7,8], and information filtering[9,10].

In this paper, an intelligent content-based filtering that can automatically and intelligently filter Web documents based on the user preferences by utilizing topic identification is proposed. Our topic identification process is based on a classification method which uses a supervised machine learning approach to classify the documents into a predefined set of categories. Web documents tend to have unpredictable characteristics, i.e., differences in length, quality and authorship. Motivated by these fuzzy characteristics, the fuzzy association concept in classifying Web documents into a predefined set of categories is adopted in our approach. Fuzzy association uses a concept of *fuzzy set* theory[11] to model the vagueness in the information retrieval process. The basic concept of fuzzy association involves the construction of a pseudothesaurus of keywords or index terms from a set of documents[12]. By constructing a pseudothesaurus, the relationship among different index terms or keywords in the documents is captured, i.e., each pair of words has an associated value to distinguish itself from other pairs of words. Therefore, the ambiguity in word usage is minimized.

Several researches have been done in the area of Web document classification. Most of these researches perform experiments using only a document set from a single Web site. However, the process of organizing the Web directories is based on human efforts and can be very subjective. Therefore, we apply our approach and perform the experiments using data sets collected from three different Web directories: *Yahoo!*[13], *Open Directory Project*[14], and *Excite*[15]. These human-compiled directories are used as the domain knowledge for topic identification and the category names are used as the topics for the documents.

The rest of the paper is organized as follows. In the next section, our proposed classification model for topic identification is described. In Section 3, the experimental setups and data sets are described. In Section 4, the results and discussions are given. The paper is concluded in Section 5.

2. Fuzzy Association Learning for Document Classification

2.1. *Fuzzy Association in Information Retrieval*

Fuzzy set theory[11] deals with the representation of classes whose boundaries are not well defined. The key idea is to associate a membership function with the elements of the class. This function takes values on the interval [0,1] with 0 corresponding to no membership in the class and 1 corresponding to full membership.

Fuzzy associative information retrieval (IR) mechanism is formalized within the fuzzy set theory and based on the definition of fuzzy association. It captures the association between the keywords to improve the retrieval results from the traditional IR System. By providing the association between the keywords, some additional documents that are not directly indexed by the keywords in the query can also be retrieved. The construction of the association between index terms or keywords is generally known as the generation of the fuzzy pseudothesaurus. The formal definitions and process of generating fuzzy pseudothesaurus based on co-occurrences of keywords can be summarized as follows[12].

Definition 2.1. Given a set of index terms, $T = \{t_1, \ldots, t_u\}$, and a set of documents, $D = \{d_1, \ldots, d_v\}$, each t_i is represented by a fuzzy set $h(t_i)$ of documents; $h(t_i) = \{F(t_i, d_j) \mid \forall d_j \in D\}$, where $F(t_i, d_j)$ is the significance (or membership) degree of t_i in d_j.

Definition 2.2. The fuzzy related terms (RT) relation is based on the evaluation of the co-occurrences of t_i and t_j in the set D and can be defined as follows.

$$RT(t_i, t_j) = \frac{\sum_k min(F(t_i,d_k), F(t_j,d_k))}{\sum_k max(F(t_i,d_k), F(t_j,d_k))}$$

A simplification of the fuzzy RT relation based on the co-occurrence of keywords[16] is given as follow.

$$r_{i,j} = \frac{n_{i,j}}{n_i + n_j - n_{i,j}}, \qquad (1)$$

where $r_{i,j}$ represents the fuzzy RT relation between keyword i and j, $n_{i,j}$ is the number of documents containing both i^{th} and j^{th} keywords, n_i is the number of documents including the i^{th} keyword, and n_j is the number of documents including the j^{th} keyword.

2.2. *Fuzzy Classification Model*

The process of classifying Web documents in our approach is explained in details as follows. Given $C = \{C_1, C_2, \ldots, C_m\}$, a set of categories, where m is the number of categories, the first step is to collect the training sets of Web documents, $TD = \{TD_1, TD_2, \ldots, TD_m\}$, from each category in C. This step involves crawling through the hypertext links encapsulated in each document. Next, the documents are cleaned through the stemming and stopword removal process, and the keywords from TD are extracted and put into separate keyword sets, $CK = \{CK_1, CK_2, \ldots, CK_m\}$. The *document frequency-inverse category frequency (df-icf)* strategy, adapted from the *tf-idf* concept[17], is proposed to select and rank the keywords within each category based on the number of documents in which the keyword appears (i.e., df) and the inverse of the number of categories in which the keyword appears (i.e., icf).

$$df\_icf(k, C_i) = DF(k, C_i) \times ICF(k), \qquad (2)$$

where $DF(k, C_i)$ is the number of documents in which keyword k occurs at least once, $ICF(k) = log(\frac{|C|}{CF(k)})$, $\mid C \mid$ is the total number of categories, and $CF(k)$ is the number of categories in which the keyword k occurs at least once.

Let $A = \{k_1, k_2, \ldots, k_n\}$ be the set of all distinct keywords from CK, where n is the number of all keywords. Then, the keyword correlation matrix M is generated via Eq.(1). The M matrix is an $n \times n$ symmetric matrix whose element $r_{i,j}$ has the value on the interval [0,1] with 0 indicates no relationship and 1 indicates full relationship between two keywords k_i and k_j.

Table 1. Predefined category sets from three Web portals.

Yahoo!		ODP		Excite	
Category	Abbr.	Category	Abbr.	Category	Abbr.
Arts & Humanities	art	Arts	art	Autos	auto
Business & Economy	bus	Business	bus	Computers	com
Computers & Internet	com	Computers	com	Entertainment	et
Education	edu	Games	gm	Games	gm
Entertainment	et	Health	hl	Health	hl
Government	gov	Home	hm	Home & Real Estate	hm
Health	hl	Kids and Teens	kid	Investing	inv
News & Media	news	News	news	Lifestyle	life
Recreation & Sports	rec	Recreation	rec	Music	music
Science	sci	Science	sci	Relationships	rel
Social Science	sosci	Shopping	shop	Sports	sport
Society & Culture	soc	Society	soc	Travel	travel
		Sports	sport		
TOTAL	12	TOTAL	13	TOTAL	12

To classify a test document d into category C_i, a set of keywords from CK_i are used to represent C_i. Then d is cleaned and its set of representative keywords is extracted from A. That is, d=$\{| k_1 |, | k_2 |, \ldots, | k_n |\}$, where $| k_i |$ is the frequency that k_i appeared in d. After that, the membership degree between d and C_i is calculated using the following equation.

$$\mu_{d,C_i} = \sum_{\forall k_a \in d} [1 - \prod_{\forall k_b \in CK_i} (1 - r_{a,b})], \tag{3}$$

where μ_{d,C_i} is the membership degree of d belonging to C_i, and $r_{a,b}$ is the fuzzy relation between keyword $k_a \in d$ and keyword $k_b \in CK_i$.

Document d is classified into category C_i when μ_{d,C_i} is the maximum for all i. The keyword k_a in d is associated to category C_i if the keywords k_b in CK_i are related to k_a. Whenever there is at least one keyword in CK_i which is strongly related to k_a (i.e., $r_{a,b} \sim 1$), then Eq.(3) yields $\mu_{d,C_i} \sim 1$, and the keyword k_a is a good fuzzy index for the category C_i. In the case when all keywords in CK_i are either loosely related or unrelated to k_a, then k_a is not a good fuzzy index for C_i (i.e., $\mu_{d,C_i} \sim 0$).

3. Experiment Setup

3.1. *Experimental Data Sets*

Experiments using the predefined categories as document topics and the document sets collected from three Web portals, *Yahoo!*[13], *Open Directory Project - ODP*[14], and *Excite*[15] are conducted. In our experiments, we only consider documents in English and ignore all other non-English documents and the selected categories are shown in Table 1. Based on these predefined categories, we collected approximately 9,000 documents from each of the Web directories as the training and test data sets. To avoid the problem of over-fitting the data when performing the experiments, we randomly select two-third of the document sets as the training set and one-third as the test set.

For the *Yahoo!* training data set, 100 keywords whose df-icf values are the highest among all keywords are selected from each of its 12 categories. Next, we combine these keywords into the set of 1074 distinct keywords. Similarly, there are 1234 distinct keywords selected from the *ODP* training data set and 1140 distinct keywords selected from the *Excite* training data set.

3.2. *Vector Space Model*

The vector space model[18] is one of the classical clustering methods. This method has been successfully applied to many IR systems including the well-known SMART system[19]. The vector space model assigns the attributes (keywords in this context) into n-dimensional space, where n is the number of the keywords. Therefore, each document can be represented by an n-dimensional vector called document vector. For the classification problem, we have some predefined set of categories, where each category can also be represented by an n-dimensional vector called category vector. To construct the representation vector for each category, the well-known *term frequency-inverse document frequency (tf-idf)*[17] is used.

To classify a document into one of the categories, first the test document vector is constructed by using the term frequency. Next, the test document vector is compared with all category vectors using a similarity metrics. The document is classified into the category where the similarity measure is the highest among all other categories. Several approaches for calculating the similarity measure between documents have been proposed[20]. Two types of measures have been widely used. The first is the distance metrics (representing dissimilarity) such as Euclidean distance. The second type is the similarity measures such as cosine and dice coefficients. In this paper, as a comparison approach, the cosine coefficient is used to calculate the similarity measure between a document and a category. The calculation of the cosine coefficient is given below.

$$COSINE(\vec{f_i}, \vec{g_j}) = \frac{\sum_{k=1}^{n} (f_{i,k} \times g_{j,k})}{\sqrt{\sum_{k=1}^{n} f_{i,k}^2 \times \sum_{k=1}^{n} g_{j,k}^2}}, \tag{4}$$

where $\vec{f_i} \in F$, $\vec{g_i} \in G$, F and G are the sets of document vectors and category vectors with n dimensions respectively, and n represents the number of keywords.

4. Experimental Results and Discussions

To compare the performance of our method (*Fuzzy*) to the vector space model (*Vector*) approach, we use the test data sets from the three Web directories and measure the classification accuracy by varying the vector lengths of the category vectors, i.e., the number of category representation keywords.

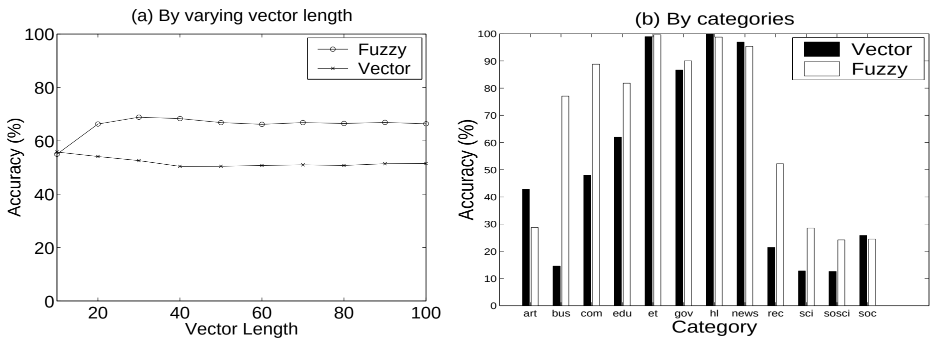

Fig. 1. Classification performance comparison - *Yahoo!*

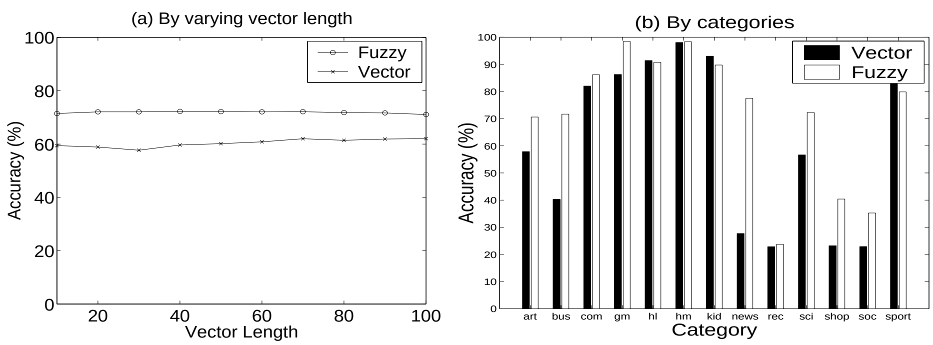

Fig. 2. Classification performance comparison - *ODP*

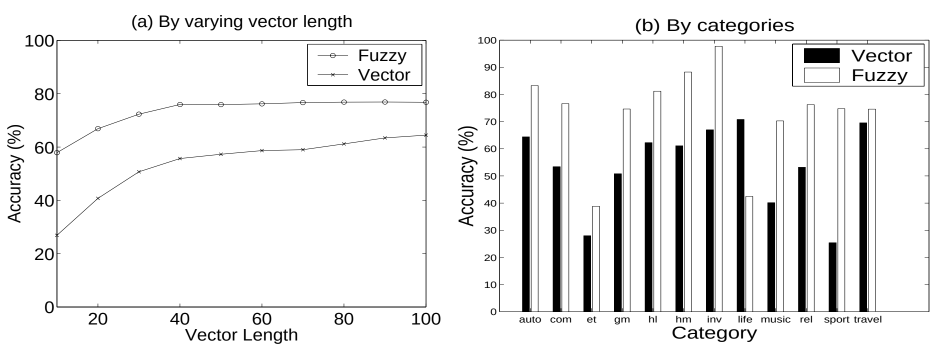

Fig. 3. Classification performance comparison - *Excite*

Fig. 1(a) and (b) show the experimental results for the *Yahoo!* data set. As can be seen from Fig. 1(a), our approach yields a higher accuracy compared to the vector space model when the vector length is increased. For example, when the vector length is equal to 100, our approach yields the accuracy of 66.4%; whereas the vector model yields the accuracy of 51.5%. In Fig. 1(b), the performance result based on the 12 categories of *Yahoo!* is presented. As expected, our approach yields higher accuracies for most of the categories.

We perform the same experiments on the *ODP* and *Excite* data sets and the experimental results are shown in Fig. 2(a) and (b) for the *ODP* data set, and in Fig. 3(a) and (b) for the *Excite* data set. The results are similar to the results obtained from the *Yahoo!* data set, except one different observation. For the *Excite* data set, the classification accuracies of both the *Vector* and *Fuzzy* methods are more sensitive to the vector length increment. As can be seen from Fig. 3(a), when the vector length is between 10 to 40, the accuracy for the *Fuzzy* method gradually increases from 57.9% to 75.9%, and the accuracy for the *Vector* method increases from 26.9% to 55.7%. The sensitivity to the number of category representation keywords is varied depending on the characteristics of the data set. For the *Excite* data set, each test document is likely to contain those keywords that belong to multiple categories. Therefore, increasing the number of keywords in the category representation vector helps improving the accuracy as more keywords are used to identify the category. For the *ODP* data set, the classification accuracy of both the *Vector* and *Fuzzy* methods are very stable through the increase of the vector length. Based on this observation, the classification model for the *ODP* data set can be minimized without losing much accuracy by using only a small number of keywords for its category representations.

5. Conclusion

In this paper, a fuzzy classification approach that automatically identifies topics for Web documents via a classification technique was proposed. Our approach adopts the fuzzy association concept as a machine learning technique to classify the documents into some predefined categories or topics. Realizing the ambiguity in word usage in English, the fuzzy association learning method avoids this problem by capturing the relationship or association among different index terms or keywords in the documents. The result is that each pair of words has an associated value to distinguish itself from other pairs of words. We performed several experiments using the data sets obtained from three different Web directories: *Yahoo!*, *Open Directory Project* and *Excite*. We compared our approach to the vector space model approach. The results show that, our approach yields higher classification accuracies compared to the vector space model when varying the number of category representation keywords. In addition, our approach is shown to work well for Web documents whose contents are highly varied in length, quality, and authorship.

Acknowledgment

S.-C. Chen gratefully acknowledges the support received from the National Science Foundation through grant CDA-9711582 at Florida International University.

References

1. J. Wang, A Survey of Web Caching Schemes for the Internet, *ACM Comp. Comm. Review* **29(5)** (1999) 36–46.
2. R. Cooley, B. Mobasher and J. Srivastava, Web Mining: Information and Pattern Discovery on the World Wide Web, *Proc. 9th IEEE Int. Conf. on Tools with Artificial Intelligence (ICTAI'97)*, Newport Beach, CA, November 1997, 558–567.
3. J. Pitkow and P. Pirolli, Mining Longest Repeating Subsequences to Predict World Wide Web Surfing, *Proc. USENIX Symp. on Internet Tech. and Syst. (USITS'99)*, Boulder, CO, October 1999, 139–150.
4. M.-L. Shyu, S.-C. Chen, and C. Haruechaiyasak, Mining User Access Behavior on the WWW, *IEEE Int. Conf. on Syst., Man, and Cybernetics*, Tucson, AZ, October 2001, 1717–1722.
5. A. Z. Broder, S. C. Glassman, M. S. Manasse, and G. Zweig, Syntactic Clustering of the Web, *Proc. 6th Int. World Wide Web Conf. (WWW6)*, Santa Clara, CA, April 1997, 391–404.
6. M.-L. Shyu, S.-C. Chen, C. Haruechaiyasak, C.-M. Shu, and S.-T. Li, Disjoint Web Document Clustering and Management in Electronic Commerce, *7th Int. Conf. on Dist. Mult. Syst. (DMS2001)*, Taipei, Taiwan, September 2001, 494–497.
7. S. T. Dumais and H. Chen, Hierarchical Classification of Web Content, *Proc. 23rd Int. ACM Conf. on Research and Development in Information Retrieval (SIGIR)*, Athens, Greece, August 2000, 256–263.
8. S. Tiun, R. Abdullah and T. E. Kong, Automatic Topic Identification Using Ontology Hierarchy, *Proc. of 2nd Int. Conf. Comput. Linguistics and Intelligent Text Processing (CICLing 2001)*, February 2001, 444–453.
9. H. Lieberman, N. W. Van Dyke and A. S. Vivacqua, Let's Browse: A Collaborative Browsing Agent, *Knowledge-Based Syst.* **12(8)** (1999) 427–431.
10. P. Resnick, N. Iacovou, M. Sushak, P. Bergstrom, and J. Riedl, Grouplens: An Open Architecture for Collaborative Filtering of Netnews, *Proc. of ACM 1994 Conf. on Comput. Supported Cooperative Work*, Chapel Hill, NC, October 1994, 175–186.
11. L. A. Zadeh, Fuzzy sets, *Inform. Contr.* **8** (1965) 338–353.
12. S. Miyamoto, T. Miyake and K.Nakayama, Generation of a Pseudothesaurus for Information Retrieval Based on Cooccurences and Fuzzy Set Operations, *IEEE Trans. on Syst., Man, and Cybernetics* **13(1)** (1983) 62–70.
13. http://www.yahoo.com, *Yahoo! Web Search Directory.*
14. http://dmoz.org, *Open Directory Project - ODP.*
15. http://www.excite.com, *Excite Directory.*
16. Y. Ogawa, T. Morita, and K. Kobayashi, A Fuzzy Document Retrieval System Using the Keyword Connection Matrix and a Learning Method, *Fuzzy Sets and Systems*, **39** (1991) 163–179.
17. G. Salton and C. Buckley, Term-Weighting Approaches in Automatic Text Retrieval, Info. Proc. Mgt. **24(5)** (1988) 513–523.
18. G. Salton, A. Wong and C. S. Yang, A Vector-Space Model for Information Retrieval, *Comm. of the ACM* **18(11)** (1975) 613–620.
19. G. Salton, ed. *The SMART retrieval system: experiments in automatic document processing*, (Prentice-Hall Series in Automatic Computation, Englewood Cliffs, New Jersey, 1971).
20. E. Rasmussen, Chapter 16: Clustering Algorithms, in *Information Retrieval: Data Structures & Algorithms*, eds. W. B. Frakes and R. Baeza-Yates (Prentice Hall, 1992) 419–442.