Web Document Classification Based on Fuzzy Association

Choochart Haruechaiyasak, Mei-Ling Shyu
Department of Electrical and Computer Engineering
University of Miami
Coral Gables, FL 33124, USA
charuech@miami.edu, shyu@miami.edu

Shu-Ching Chen
Distributed Multimedia Information System Laboratory
School of Computer Science
Florida International University
Miami, FL 33199, USA
chens@cs.fiu.edu

Xiuqi Li
NSF/FAU Multimedia Laboratory, Florida Atlantic University
Boca Raton, FL 33431, USA
xli@cse.fau.edu

Abstract

In this paper, a method of automatically classifying Web documents into a set of categories using the fuzzy association concept is proposed. Using the same word or vocabulary to describe different entities creates ambiguity, especially in the Web environment where the user population is large. To solve this problem, fuzzy association is used to capture the relationships among different index terms or keywords in the documents, i.e., each pair of words has an associated value to distinguish itself from the others. Therefore, the ambiguity in word usage is avoided. Experiments using data sets collected from two Web portals: Yahoo! (www.yahoo.com) and Open Directory Project (dmoz.org) are conducted. We compare our approach to the vector space model with the cosine coefficient. The results show that our approach yields higher accuracy compared to the vector space model.

Keywords: Information Processing on the Web, Data Mining, Document Classification, Fuzzy Association.

1. Introduction

The World Wide Web (WWW) can be viewed as a distributed database system, but with two different aspects. Firstly, WWW contains much larger amount of data than a typical database system. WWW is often referred to as the world's largest distributed database system with the amount of data growing at an exponential rate [17]. These data can be of heterogeneous types such as text, image, audio, and video. Secondly, WWW involves a huge user population that is not restricted to a certain demographic group or a geographic area. The result is the wide variation in information content and quality. In addition, unlike a typical database system where the majority of users only retrieve the information through some queries, WWW allows its users to provide and share the information publicly on the system. With the large amount of available information on the Web, searching for specific information or discovering any useful information becomes a difficult and challenging task. To alleviate this problem, many data mining techniques have been applied into the Web context. This is referred to as *Web mining* [3]. Web mining is defined as the discovery and analysis of useful information from WWW. Some of Web mining techniques include analysis of user access patterns [10][14], Web document clustering [1][15], and classification [2][4][5][16].

Document classification or text categorization (as used in information retrieval context) is the process of assigning a document to a predefined set of categories based on the document content. Document classification can be applied as an information filtering tool and can also be used to improve the retrieval results from a query process. To help the users search and browse for specific information on the Web, many of the well-known Web portals such as *Yahoo!* [21] have organized the information, in form of Web documents, into some predefined categories such as *Arts & Humanities*, *Computers & Internet*, and *Entertainment*. However, this

approach of organizing Web documents requires human efforts and hence, is very subjective and does not scale well.

In this paper, a method of automatically classifying Web documents into a set of categories using the *fuzzy association* concept is proposed. The fuzzy association uses the concept of the *Fuzzy Set* theory [18] to model the vagueness in the information retrieval process. Examples of the research works involving the use of the fuzzy association technique include [6], [7], [8], and [9]. The basic concept of fuzzy association involves the construction of a *pseudothesaurus* of keywords or index terms from a set of documents [7]. 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 document classification or text categorization. Some of these researches perform experiments using only a document set from a specific topic. For example, in [5], the document collection, *Reuters*, which is business related, is used in their experiments. Other research work such as [2], [4], and [16] focus on the Web documents. However, all of these researches use only a set of documents obtained from a single Web directory. For example, [2] and [16] use Yahoo! Directory as their data set, and [4] uses *LookSmart* (*www.looksmart.com*)'s directory. As mentioned earlier, the process of organizing the Web directories is based on human efforts and can be very subjective. Therefore, in this paper, we apply our approach and perform our experiments using data sets collected from two different Web portals: *Yahoo!* [21] and *Open Directory Project* [19].

In general, when dealing with data in high multi-dimensional space, the performance, in terms of storage space and execution time, can be greatly affected by the high dimension. This problem is generally known as *the curse of dimensionality*. For a document data set, this problem also holds, since a document collection can contain millions of different index terms or keywords. A classical document clustering approach, *vector space* model [13], which represents each document using *n*-dimensional vector (where *n* is the number of keywords) also suffers from this problem. By using the fuzzy association technique in our approach, the dimension of the keyword representation for the categories can be reduced without much performance degradation. Also the selection of different keywords in representing each category does not affect the performance as much compared to the vector space approach.

The rest of the paper is organized as follows. In the next section, the concept of the fuzzy association that has been applied in the area of information retrieval systems is introduced. In this section, our proposed fuzzy classification model is also described. In Section 3, the experimental results and discussions are given. The paper is concluded in Section 4.

2. Fuzzy association for document classification

In this section, we first review the concept of the fuzzy association that has been applied in the area of information retrieval systems. Then we describe our classification model based on the fuzzy association concept in details.

2.1. Fuzzy association in information retrieval

Fuzzy set theory [18] 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. Membership values between *0* and *1* indicate *marginal* elements of the class. Thus, membership in a fuzzy set is a notion intrinsically *gradual* instead of abrupt or *crisp* (as in conventional Boolean logic).

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 traditional IR systems. 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.

Definition 1. *A fuzzy association between two finite sets* $X=\{x_1,\ldots,x_u\}$ *and* $Y=\{y_1,\ldots,y_v\}$ *is formally defined as a binary fuzzy relation* $f: X \times Y \rightarrow [0,1]$, *where u and v are the numbers of elements in X and Y, respectively.*

The construction of the association between index terms or keywords is generally known as the generation of the fuzzy pseudothesaurus. In [7], a formal definition and process of generating fuzzy pseudothesaurus based on co-occurrences of keywords is given. It can be summarized as follows.

Definition 2. *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;* $h(t_i)=\{F(t_i,d_j) \mid \forall\ d_i \in D\}$, *where* $F(t_i,d_j)$ *is the significance (or membership) degree of* t_i *in* d_j.

Definition 3. *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.*

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

In [9], a simplification of the fuzzy *RT* relation based on the co-occurrence of keywords is given as follow.

$$r_{i,j} = \frac{n_{i,j}}{n_i + n_j - n_{i,j}}, \qquad \text{Eq. 1}$$

where

- $r_{i,j}$ represents the fuzzy *RT* relation between keywords *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.

Next, the calculation of the fuzzy *RT* relation between keywords is applied in our classification model.

2.2. Fuzzy classification model

The process of classifying Web documents is explained in details as follows. Given $C = \{C_1, C_2, ..., C_m\}$, a set of categories, where m is the total number of categories, the first step is to collect the training sets of Web documents, $TD = \{TD_1, TD_2, ..., TD_m\}$, from each category in *C*. This step involves *crawling* through the hypertext links encapsulated in each document. Once the document collections are obtained, they are cleaned through the stemming and stopword removal processes. Next, the most frequently occurred keywords from the document sets based on each category are extracted and put into separate keyword sets, $K = \{K_1, K_2, ..., K_m\}$. From these m sets of keywords, we combined them into a set of all keywords, $A = \{k_1, k_2, ..., k_n\}$,where n is the total number of all distinct keywords representing the vector dimension. Note that some of the keywords can appear in more than one category, but we only consider one instance of these. Then we generate the keyword correlation matrix *M* using the fuzzy *RT* relation equation (given in Eq. 1). The keyword correlation matrix is an $n \times n$ symmetric matrix whose element, m_{ij}, has the value on the interval [*0*, *1*] with *0* indicates no relationship and *1* indicates full relationship between the keywords k_i and k_j. Therefore, m_{ij} is equal to 1 for all $i = j$, since a keyword has the strongest relationship to itself.

To classify the documents in the test data set into different categories, first, each category must be represented with a set of keywords. The best way to represent each category is to select only the exclusive keywords, i.e., for category C_i, we consider the keywords in K_i which do not belong in another keyword sets K_j, where $j=1...m$ and $j \neq i$. We refer to this as the *category keyword sets*, $CK = \{CK_1, CK_2, ..., CK_m\}$. Next, the test documents in the test data set are cleaned and the keywords are extracted by looking up in *A,* the list of all keywords. This process gives us the representation of those test documents, $D = \{d_1, d_2, ..., d_p\}$, where p is the total number of documents to be classified. After that, the membership degrees between each document to each of the category sets are calculated using the following equation.

$$\mu_{i,j} = \max_{\forall k_a \in d_i} [1 - \prod_{\forall k_b \in CK_j} (1 - r_{a,b})], \qquad \text{Eq. 2}$$

where

- $\mu_{i,j}$ is the membership degree of d_i belonging to C_j,
- $r_{a,b}$ is the fuzzy relation between keyword $k_a \in d_i$ and keyword $k_b \in CK_j$.

A document d_i is classified into the category C_j where the membership degree $\mu_{i,j}$ is the maximum. The keyword k_a in d_i is associated to category C_j if the keywords k_b*'s* in CK_j (for category C_j) are related to the keyword k_a. Whenever there is at least one keyword in CK_j which is strongly related to the keyword k_a in d_i (i.e., $r_{a,b} \sim 1$), then Eq. 2 yields $\mu_{i,j} \sim 1$, and the keyword k_a is a good fuzzy index for the category C_j. In the case when all keywords in CK_j are either loosely related or unrelated to k_a, the keyword k_a is not a good fuzzy index for C_j (i.e., $\mu_{i,j} \sim 0$).

3. Experiments and results

This section provides the descriptions and characteristics of the data sets used for performing our experiments. Also, we briefly review the *vector space model* with the *cosine coefficient* as a comparison approach. Then, the experimental results and discussions are presented.

3.1. Experimental data sets

Experiments using the predefined categories and the document sets collected from two Web portals, *Yahoo!* [21] and *Open Directory Project (ODP)* [19], are conducted. The brief description and history of these two Web portals are provided in [20]. In our experiments, we only consider those documents in English and ignore all other non-English documents. Therefore, the categories, *World* and *Regional*, are excluded from our experimental

data sets. Table 1 shows the selected categories from these two Web portals. Based on these predefined categories, we collect approximately *18,000* 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 documents as the training data set and one-third as the test data set.

Table 1. Predefined category sets from two Web portals

Yahoo!		***ODP***	
Category	**Abbr.**	**Category**	**Abbr.**
Arts & Humanities	art	Arts	art
Business & Economy	bus	Business	bus
Computers & Internet	com	Computers	com
Education	edu	Games	game
Entertainment	et	Health	health
Government	gov	Home	home
Health	health	Kids and Teens	kid
News & Media	news	News	news
Recreation & Sports	rec	Recreation	rec
Science	sci	Science	sci
Social Science	sosci	Shopping	shop
Society & Culture	soc	Society	soc
TOTAL	**12**	Sports	sport
		TOTAL	**13**

Considering only the training data sets from these two different Web sites, we extract and select the most frequently occurred keywords from each category as follows. For the *Yahoo!* data set, *350* most frequent keywords are selected from each of *12* categories. Some of the keywords appear in more than one category, but we only consider one instance for each of these. The total number of all distinct keywords is *2033*. For the *ODP* data set, we also select *350* most frequent keywords from each of *13* categories. The total number of distinct keywords is *1889*.

3.2. Vector space model

The *vector space model* is one of the classical clustering methods first proposed by [13]. This method has been successfully applied to many IR systems including the well-known SMART system [12]. The *vector space model* assigns the attributes (keywords in this context) into *n*-dimensional space, where *n* is the number of the attributes. Therefore, each document can be represented by an *n*-dimensional vector called a *document vector*. For the classification problem, we have some predefined set of categories, where each can also be represented by an *n*-dimensional vector called *category vector*. To classify a document into one of the categories, the document vector is compared with all category vectors using a similarity metric. 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 [11]. Two types of measures have been widely used. The first is the distance metric (representing dissimilarity) such as *Euclidean distance*. The second type is similarity measures such as *cosine* and *dice coefficients*. In this paper, as a comparison approach, the *cosine coefficient* is used to calculate the similarity measures between a document and a category. The calculation of the *cosine coefficient* is given below.

$$COSINE(\vec{f}_i, \vec{g}_j) = \frac{\sum_{k=1}^{n}(f_{i,k} \times g_{j,k})}{\sqrt{\sum_{k=1}^{n} f_{i,k}^{\ 2} \times \sum_{k=1}^{n} g_{j,k}^{\ 2}}} \qquad \text{Eq. 3}$$

where

- $\vec{f}_i \in F$, F is a set of *n*-dimensional document vectors,
- $\vec{g}_j \in G$, G is a set of *n*-dimensional category vectors, and
- *n* represents the total number of distinct keywords.

3.3. Results and discussions

To compare the performance of our method (denoted as *Fuzzy*) to the vector space model (denoted as *Vector)* approach, we use the test data sets and measure the classification accuracy by varying the vector lengths of the category vectors. To see the effect of using different sets of keywords in representing the category vectors, we provide two ways of selecting the keywords: selecting from the most frequently occurred keywords (denoted as *topmost)*, and selecting from the least frequently occurred keywords (denoted as *bottommost*).

Figure 1 shows the experimental result by using the *Yahoo!* data set. As can be seen from this figure, for all cases, the classification accuracy increases when the number of keywords used to represent the category vectors is increased. Our approach yields a higher accuracy compared to the *vector space model*. For example, when the vector length is *10*, our approach yields the accuracies of 74.9% for the *topmost* sets and 41.1% for the *bottommost* sets, whereas the *vector space model* yields the accuracies of 57.0% for the *topmost* sets and 12.2% for the *bottommost* sets. In Figure 2, the performance result based on *12* categories of *Yahoo!* is presented. As expected, our approach yields higher accuracies for all categories.

We perform the same experiments on the *ODP* data set. The results are shown in Figure 3 and Figure 4, respectively. The results are similar to the results obtained from the *Yahoo!* data set, except one different observation. By using the *bottommost* keywords in our approach, the average accuracy is 78.1%, and by using the *topmost* keywords in the *vector space model*, the average accuracy is 67.1%. That is, by using either the *topmost* or *bottommost* representations, our approach performs better than the *vector space model*.

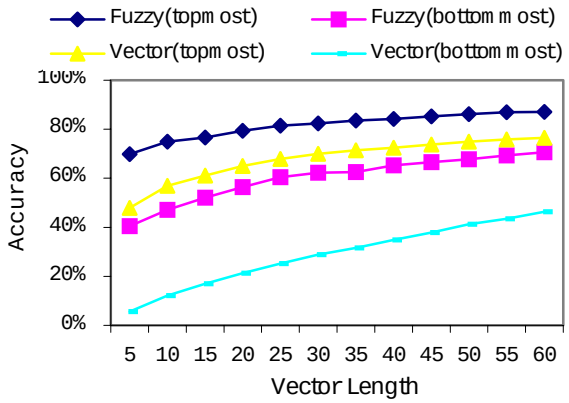

Figure 1. Classification performance by varying the vector length - Yahoo!

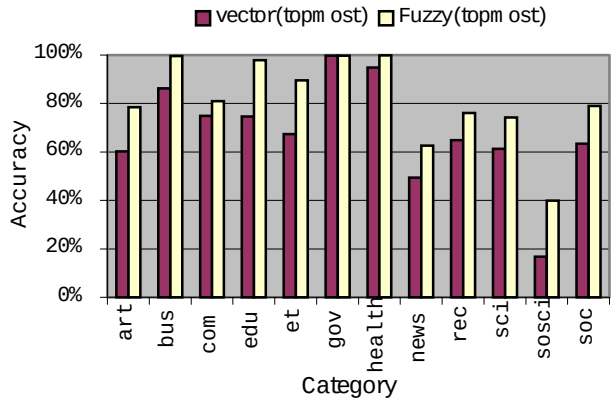

Figure 2. Classification performance by categories - Yahoo!

Table 2 shows the summarized results for both *Yahoo!* and *ODP* data sets. The results are calculated by averaging the accuracy values over all the vector lengths. By using the *topmost* representation for the category vector, our approach yields higher average classification accuracies of 13.7% and 17.7% over the *vector space model* for the *Yahoo!* and *ODP* data sets, respectively. Another observation is that, for our approach, the selection of different keywords in representing the category does not affect the performance as much as the *vector space model*. For example, for the *Yahoo!* data set, by using the *bottommost* keywords, instead of the *topmost* keywords, the accuracy drops 21.4% in our approach, whereas the accuracy drops 39.0% in the *vector space model* approach.

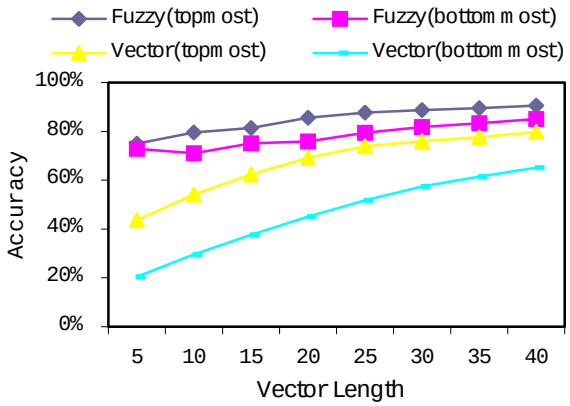

Figure 3. Classification performance by varying the vector length - ODP

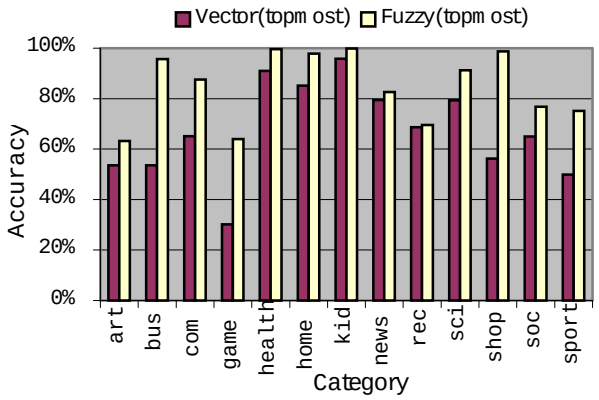

Figure 4. Classification performance by categories - ODP

Table 2. Average classification accuracy

Data set	*Fuzzy* (*topmost*)	*Fuzzy* (*bottommost*)	*Vector* (*topmost*)	*Vector* (*bottommost*)
Yahoo!	81.5%	60.1%	67.8%	28.8%
ODP	84.8%	78.1%	67.1%	46.1%

4. Conclusion

In this paper, an alternative approach of automatically classifying the Web documents into some predefined categories using the fuzzy association concept is proposed. Realizing the ambiguity in word usage in English, the fuzzy association 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. Experiments using the data sets obtained from two different Web directories, *Yahoo!* and *ODP*, are conducted. Both Web portals are independent and have different characteristics from each other. We compare our fuzzy association approach to the *vector space model* approach. To see the effect of different keyword selections for category vectors, two different alternatives: selecting from the most frequently occurred keyword (*topmost*) and selecting from the least frequently occurred keywords (*bottommost*) with varying vector lengths are used. The results show that, on average, our approach yields higher classification accuracies compared to the *vector space model* for both the *topmost* and *bottommost* cases. In addition, with our approach, using fewer numbers of keywords for category representation does not degrade the accuracy as much compared with the *vector space model*.

5. Acknowledgments

For Shu-ching Chen, this research was supported in part by NSF CDA-9711582.

6. References

[1] A.Z. Broder, S.C. Glassman, and M.S. Manasse, "Syntactic Clustering of the Web," *Proceedings of the 6th International World Wide Web Conference*, April 1997, pp. 391-404.

[2] C. Chekuri, M. Goldwasser, P. Raghavan, and E. Upfal, "Web Search Using Automatic Classification," *Proceedings of the 6th International World Wide Web Conference*, April 1997.

[3] R. Cooley, B. Mobasher, and J. Srivastava, "Web Mining: Information and Pattern Discovery on the World Wide Web," *Proceedings of the 9th IEEE International Conference on Tools with Artificial Intelligence (ICTAI'97)*, November 1997, pp. 558-567.

[4] S. T. Dumais and H. Chen, "Hierarchical Classification of Web Content," *Proceedings of the 23rd International ACM Conference on Research and Development in Information Retrieval (SIGIR'00)*, August 2000, pp. 256-263.

[5] D. Koller and M. Sahami, "Hierarchically Classifying Documents Using Very Few Words," *Proceedings of the 14th International Conference on Machine Learning (ICML'97)*, July 1997, pp. 170-178.

[6] S. Miyamoto, "Two Approaches for Information Retrieval Through Fuzzy Associations," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 19, no. 1, January/February 1989, pp. 123-130.

[7] S. Miyamoto, T. Miyake, and K. Nakayama, "Generation of a Pseudothesaurus for Information Retrieval Based on Cooccurences and Fuzzy Set Operations," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 13, no. 1, 1983, pp. 62-70.

[8] S. Miyamoto and K. Nakayama, "Fuzzy Information Retrieval Based on a Fuzzy Pseudothesaurus," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 16, no. 2, March/April 1986, pp. 278-282.

[9] 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*, vol. 39, 1991, pp. 163-179.

[10] J. Pitkow and P. Pirolli, "Mining Longest Repeating Subsequences to Predict World Wide Web Surfing," *Proceedings of the 2nd USENIX Symposium on Internet Technologies and Systems* (*USITS'99*), Oct 1999, pp.139-150.

[11] E. Rasmussen, "Chapter 16: Clustering Algorithms," in W. B. Frakes and R. Baeza-Yates, editors, *Information Retrieval: Data Structures &Algorithms*, Prentice Hall, 1992, pp. 419-442.

[12] G. Salton, editor. "The SMART retrieval system: experiments in automatic document processing," Prentice-Hall Series in *Automatic Computation*, Englewood Cliffs, New Jersey, 1971, Chapters 14-17.

[13] G. Salton, A. Wong, and C.S. Yang, "A Vector-Space Model for Information Retrieval," *Communications of the ACM*, vol. 18, no. 11, 1975, pp. 613-620.

[14] M.-L. Shyu, S.-C. Chen, and C. Haruechaiyasak, "Mining User Access Behavior on the WWW," *IEEE International Conference on Systems, Man, and Cybernetics*, October 2001, pp. 1717-1722.

[15] M.-L. Shyu, S.-C. Chen, C. Haruechaiyasak, C.-M. Shu, and S.-T. Li, "Disjoint Web Document Clustering and Management in Electronic Commerce," *Proceedings of the Seventh International Conference on Distributed Multimedia Systems* (*DMS'01*), September 2001.

[16] S. Tiun, R. Abdullah, and T.E. Kong, "Automatic Topic Identification Using Ontology Hierarchy," *Proceedings of the Second International Conference on Computational Linguistics and Intelligent Text Processing (CICLing'01)*, February 2001, pp. 444-453.

[17] J. Wang, "A Survey of Web Caching Schemes for the Internet," *ACM Computer Communication Review*, October 1999, pp. 36-46.

[18] L.A. Zadeh, "Fuzzy Sets," in D. Dubois, H. Prade, and R.R. Yager, editors, *Readings in Fuzzy Sets for Intelligent Systems*, Morgan Kaufmann Publishers, 1993.

[19] http://dmoz.org, *Open Directory Project – ODP*.

[20] http://www.searchenginewatch.com/links/major.html, *The Major Search Engines*.

[21] http://www.yahoo.com, *Yahoo! Web Search Directory*.