

Developing a Fuzzy Search Engine Based on Fuzzy Ontology and Semantic Search

Lien-Fu Lai Chao-Chin Wu Pei-Ying Lin

Dept. of Computer Science and Information Engineering
National Changhua University of Education
Changhua, R.O.C.
{lflai,ccwu}@cc.ncue.edu.tw, ester6236@gmail.com

Liang-Tsung Huang

Department of Biotechnology
Mingdao University
Changhua, R.O.C.
larry@mdu.edu.tw

Abstract—Most of existing search engines retrieve web pages by means of finding exact keywords. Traditional keyword-based search engines suffer several problems. First, synonyms and terms similar to keywords are not taken into consideration to search web pages. Users may need to input several similar keywords individually to complete a search. Second, traditional search engines treat all keywords as the same importance and cannot differentiate the importance of one keyword from that of another. Third, traditional search engines lack an applicable classification mechanism to reduce the search space and improve the search results. In this paper, we develop a fuzzy search engine, called Fuzzy-Go. First, a fuzzy ontology is constructed by using fuzzy logic to capture the similarities of terms in the ontology, which offering appropriate semantic distances between terms to accomplish the semantic search of keywords. The Fuzzy-Go search engine can thus automatically retrieve web pages that contain synonyms or terms similar to keywords. Second, users can input multiple keywords with different degrees of importance based on their needs. The totally satisfactory degree of keywords can be aggregated based on their degrees of importance and degrees of satisfaction. Third, the domain classification of web pages offers users to select the appropriate domain for searching web pages, which excludes web pages in the inappropriate domains to reduce the search space and to improve the search results.

Keywords- Fuzzy Search Engine; Fuzzy Ontology; Semantic Search

I. INTRODUCTION

As the massive data on the internet is increasing rapidly, internet search engines have become the essential ways to find information. Most of existing search engines, such as Google, Yahoo, MSN, ASK, Baidu, and Bing, retrieve web pages by means of finding exact keywords. These keyword-based search engines collect and analyze web pages through web crawlers. While users input keywords to search web pages, web pages that contain exact keywords are retrieved and ranked. For example, the Google search engine sorts the search results by their Page-rank scores, Relevance scores and Local scores. However, traditional keyword-based search engines suffer several problems:

- Synonyms and terms similar to keywords are not taken into consideration to search web pages. Users may need to input several similar keywords individually to

complete a search. The restriction of exact keywords makes it inconvenient for users to search web pages. Many valuable web pages would be omitted if users did not search for several similar keywords individually.

- While users input several keywords to search web pages, different keywords may have different degrees of importance in their opinions. Traditional search engines treat all keywords as the same importance and cannot differentiate the importance of one keyword from that of another.
- The problem of information overload makes it difficult for users to find really useful information from a large amount of search results. Traditional search engines lack an applicable classification mechanism to reduce the search space and improve the search results.

To alleviate the mentioned problems, we have applied the fuzzy logic theory and the semantic search techniques to develop a fuzzy search engine, called Fuzzy-Go. First, a fuzzy ontology is constructed by using fuzzy logic [15] to capture the similarities of terms in the ontology, which offering appropriate semantic distances between terms to accomplish the semantic search of keywords. The Fuzzy-Go search engine can thus automatically search web pages that contain synonyms or terms similar to keywords. Second, users can input multiple keywords with different degrees of importance based on their needs. The totally satisfactory degree of keywords can be aggregated based on their degrees of importance and degrees of satisfaction. Third, the domain classification of web pages offer users to select the appropriate domain for searching web pages, which excludes web pages in the inappropriate domains to reduce the search space and to improve the search results.

An overview of the Fuzzy-Go search engine is shown in Figure 1. The Web Crawler is developed to gather and classify web pages. Web pages are classified and stored based on their domains. The characteristic information of web pages is recorded in the Web Document. The Ontology Manager constructs and maintains a Fuzzy Ontology. A data mining approach is applied on the Web Document to calculate the fuzzy similarity between terms in the Fuzzy Ontology. The User Interface provides users to select an appropriate domain and to input multiple keywords with different degrees of

importance based on their needs. The Fuzzy Search Mechanism would exclude web pages in the inappropriate domains to reduce the search space and to improve the search results. The keywords are expanded by the Fuzzy Ontology to find out synonyms and terms similar to keywords. The search results are ordering based on several fuzzy factors including the satisfaction degrees of keywords, the importance degrees of keywords, the relevance of domains, the page ranks [3], the page in-links [2], titles, and last modified dates.

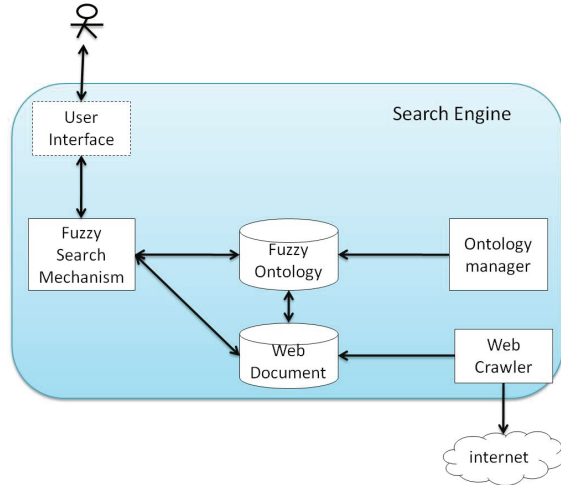


Figure 1. An overview of the Fuzzy-Go search engine

II. CONSTRUCTING THE FUZZY ONTOLOGY

Domain dependency is an important character of knowledge: the meaning of a term may be different in different domains. We construct a two-layered fuzzy ontology to organize terms that are elicited from WordNet [6]. WordNet is a large lexical database of English, built by Princeton University. Nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms, each expressing a distinct concept. In the two-layered fuzzy ontology, the first layer forms a domain hierarchy (see Figure 2). We adopt the Suggested Upper Merged Ontology (SUMO) [5] to classify domains. SUMO is owned by the IEEE and defines a group of domain ontologies which have been mapped to the WordNet domains. In our approach, the domain hierarchy is treated as gen-spec relationships. The upper level in the domain hierarchy

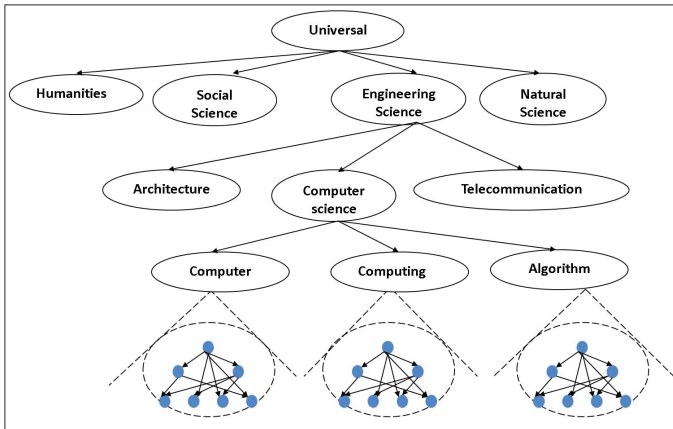


Figure 2. The domain hierarchy in the two-layered fuzzy ontology

maintains terms in a more general domain, while the lower level maintains terms in a more specific domain. Each domain contains a term lattice in the second layer (see Figure 3). A sub-domain can inherit the term lattice from its super-domain, add new terms specific to the sub-domain itself, and override (i.e., redefine) fuzzy similarity between terms in the super-domain.

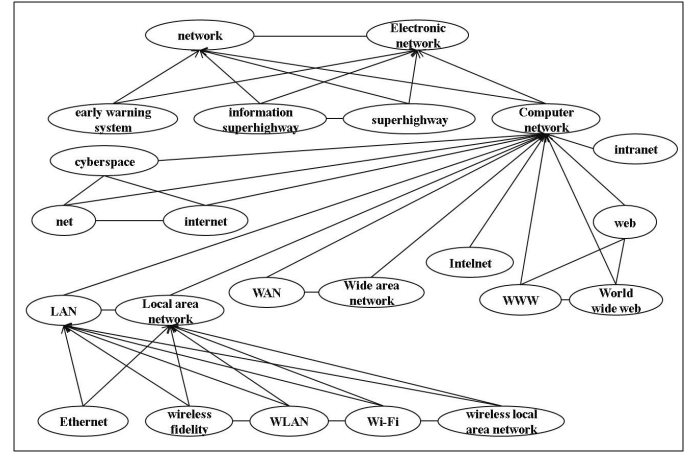


Figure 3. The term lattice in a domain

Church et al. [9] had extended the word association in psycholinguistics to provide the basis for a statistical description of the semantic relationship between words through the co-occurrence of words. To determine the similarity degree between terms, we use the data mining method to find out the degree of co-occurrence for two terms in web pages. The measure of mutual information [9] is adopted to calculate the degree of a children node conforming to a parent node in the term lattice. For any two terms x and y , the mutual information $I(x, y)$ is defined as

$$I(x, y) \equiv \log_2 \frac{P(x, y)}{P(x)P(y)} \quad (1)$$

Where $P(x)$ and $P(y)$ is the probabilities of observing x and y independently in web pages, and $P(x, y)$ is the probabilities of observing x and y together in web pages. The higher the semantic relationship between x and y , the larger the co-occurrence degree (i.e. $P(x, y)$ is larger), and consequently the larger the mutual Information $I(x, y)$.

To normalize the mutual information into a similarity degree in interval $[0, 1]$, we adopt the maximizing set [16] to assign the fuzzy similarity degree between terms. For all mutual Information $x \in X$, let f be a real-valued function in X . The maximizing set $\tilde{M} = \{(x, \mu_{\tilde{M}}(x))\}, x \in X$ assigns the degrees of membership $\mu_{\tilde{M}}(x)$ for all $x \in X$ with

$$\mu_{\tilde{M}}(x) = \frac{f(x) - \inf(f)}{\sup(f) - \inf(f)} \quad (2)$$

Where $\inf(x)$ is the minimum of f and $\sup(x)$ is the maximum of f .

The Ontology Manager is responsible to apply the data mining approach to give the similarity degrees between terms in all term lattices based on the Web Document. Figure 4 shows a term lattice with the similarity degrees. Note that the similarity degree between two synonyms defined by WordNet is assigned 1.0 directly.

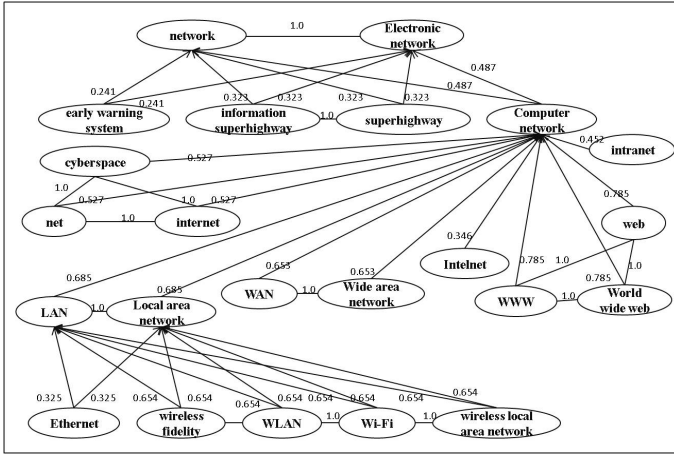


Figure 4. A term lattice with similarity degrees

III. THE WEB CRAWLER

We develop a crawler agent to automatically extract web pages and inventory them by finding out keywords and characteristic values. While the Web Crawler retrieves web pages, some important information of web pages is analyzed and stored, including titles, keywords, the page ranks of web pages, the last modified dates etc. Figure 5 shows the work flow of the Web Crawler.

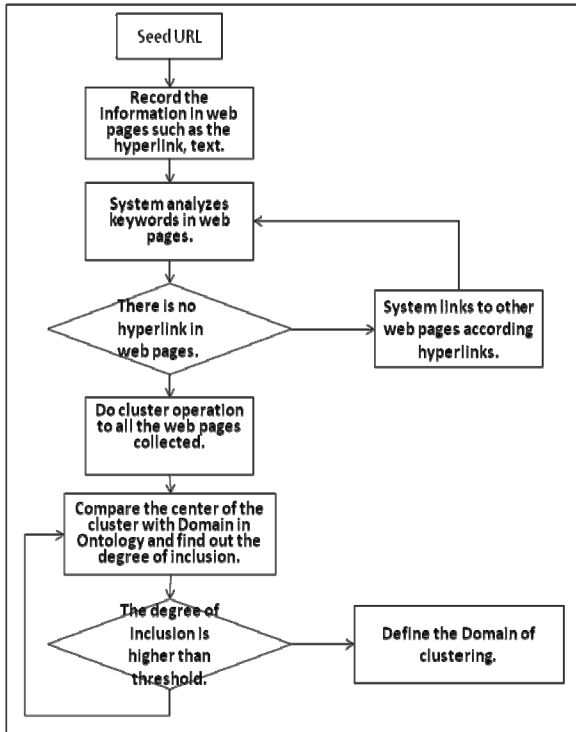


Figure 5. The work flow of the Web Crawler

The Web Crawler is also responsible to classify retrieved web pages by determining their domains. Since the amount of retrieved web pages is huge, we adopt the notions of Fuzzy C-Mean (FCM) [7] to cluster web pages based on their keywords. In our approach, each web page is translated into a vector of keywords. Each vector doesn't belong to only one cluster but belongs to several clusters with different degrees. By comparing to the terms within domains in the Fuzzy Ontology, the keywords in the center of a cluster are useful to determine what domain the cluster belongs to.

Keywords in web pages are regarded as the characteristic values and each web page is translated into a vector of keywords in the web page. Assume that all web pages in the Web Document are gathered into k clusters. The Euler distance between two web pages x_i and x_j would be

$$dist(x_i, x_j) = \left(\sum_{d=1}^k |x_{id} - x_{jd}|^2 \right)^{\frac{1}{2}} \quad (3)$$

The center of a cluster will be changed after iteration. In each iteration, we calculate the total deviation for changing the center of the cluster. In the case of k clusters, the total deviation E can be obtained by $E = \sum_{i=1}^k \sum_{x \in S_i} \|x - C_i\|^2$, where C_i

is the center of each cluster S_i and x is the web page in cluster S_i . If the total deviation for the $m+1$'s iteration is smaller than the m 's, the center of the cluster would be changed into $m+1$'s iteration -- otherwise the center would remain unchanged (see Figure 6).

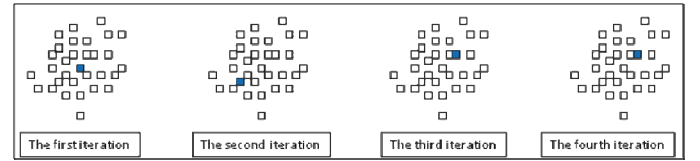


Figure 6. Changing of the FCM center of the clustering

Since the centers of clusters will be changed again and again, we use an objective function J to check if the result of clustering is stable.

$$J = \sum_{i=1}^k J_i = \sum_{i=1}^k \sum_{j=1}^n w_{ji}^m \|x_j - C_i\|^2 \quad (4)$$

Where w_{ji} is the degree of the web page x_j belongs to the cluster S_i with $w_{ji} = \frac{1}{\sum_{s=1}^k \left(\frac{\|x_j - C_i\|}{\|x_j - C_s\|} \right)^{\frac{2}{m-1}}}$, $\sum_{i=1}^k w_{ji} = 1$, w_{ji}^m forms a

weighting matrix, and C_i is the center of each cluster S_i with $C_i = \frac{\sum_{j=1}^n w_{ji}^m x_j}{\sum_{j=1}^n w_{ji}^m}$. If the value of the objective function J is stable, the result of clustering is stable.

$$Sim(C_K, C_D) = \frac{|C_K \cap C_D|}{|C_K|}, \{C_K\} \neq \emptyset \quad (5)$$

IV. THE FUZZY SEARCH MECHANISM

Domain: Engineering Science computer science please Choose or not

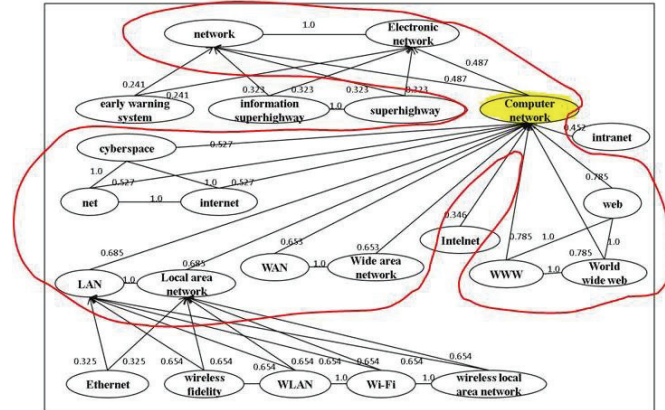
KeyWord fuzzy theory grey theory

very important most important most important very important **rather important** moderately important rather unimportant unimportant don't care

+Add KeyWord Search Advance

NCUE CSIE © 2010 All Rights Reserved

It is inconvenient for users to be restricted to input precise keywords. Users may need to input several similar keywords individually to complete a search. To alleviate the problem, we use the term expansion to find out synonyms and terms similar to keywords based on the semantics inherited from the selected domain. While the user selects an appropriate domain and inputs keywords to search web pages, the Fuzzy Search Mechanism would expand keywords to synonyms and terms similar to keywords based on the term lattice of the selected domain. For the example in Figure 9, when the keyword is ‘Computer Network’ and the α -cut is set to 0.475, we would acquire 12 terms whose similarity degrees to ‘Computer Network’ are greater than 0.475.



The fuzzy similarity between two non-adjacent terms T_A and T_B can be calculated by the multiplication principle for a path from T_A to T_B . That is, we can obtain the fuzzy similarity between T_A and T_B via another term T_C in the term lattice of the domain D .

$$Sim(T_A, T_B) = \underset{\forall T_C \in D}{Max} \{ Sim(T_A, T_C) \times Sim(T_C, T_B) \} \quad (6)$$

2687

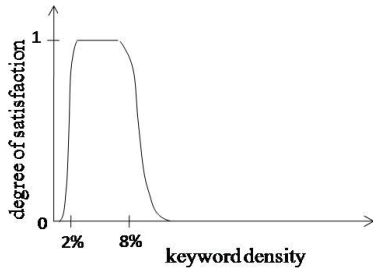


Figure 10. The membership function of the optimum keyword density

Since users may input multiple keywords with different degrees of importance to search web pages, the totally satisfactory degree of multiple keywords should be aggregated based on their degrees of importance and degrees of satisfaction. The fuzzy weighted average (FWA) [8,11] is applied to calculate the totally satisfactory degree of multiple keywords using triangular fuzzy numbers. In our approach, the degree of satisfaction between a keyword and a web page is the indicator (x_i), and the degrees of importance are weights (w_i) that act upon indicators. Therefore, the fuzzy weighted average y can be defined as

$$y = f(x_1, \dots, x_n, w_1, \dots, w_n) = \frac{\sum_{i=1}^n w_i x_i}{\sum_{i=1}^n w_i} \quad (7)$$

Where there is n keywords, the degree of satisfaction x_i , $1 \leq i \leq n$ is a crisp value, and the degree of importance w_i , $1 \leq i \leq n$ is represented by a triangular fuzzy number. We adopt the approximate expressions on $\oplus$ and $\otimes$ operators for the computation of L-R fuzzy numbers, which is suggested by Dubois and Prade [10]. For example, a user may input two keywords to search web pages: 'expert systems' with moderately important and 'soft computing' with very important. In the term lattice, the term 'expert systems' links to a similar term 'knowledge-based systems' with the similarity degree 0.786, and 'soft computing' links to 'fuzzy logic' with 0.623. Assume that the keyword densities of 'expert systems', 'knowledge-based systems', 'soft computing' and 'fuzzy logic' for the web page P are 1.5%, 1.2%, 0% and 2.1% respectively. The expanded keyword density of 'expert systems' would be $1.5\% + 1.2\% \times 0.786 = 2.4432\%$, and the degree of satisfaction between 'expert systems' and P would be 1. The expanded keyword density of 'soft computing' would be $0\% + 2.1\% \times 0.623 = 1.3083\%$, and the degree of satisfaction between 'soft computing' and P would be $s(1.3083) = 1 - 2 \left(\frac{2 - 1.3083}{2 - 0} \right)^2 \approx 0.76$.

FWA is applied to calculate P 's totally satisfactory degree of keywords as

$$y = \frac{(0.25, 0.5, 0.75) \otimes 1.0 \oplus (0.75, 1, 1) \otimes 0.76}{(0.25, 0.5, 0.75) \oplus (0.75, 1, 1)} = \frac{(0.82, 1.26, 1.51)}{(1, 1.5, 1.75)} \\ \approx \frac{1.1517}{1.3536} = 0.8508$$

Applying the mathematical operations on fuzzy numbers [10,16], we get two fuzzy numbers (0.82, 1.26, 1.51) and (1, 1.5, 1.75). The center of gravity is adopted to defuzzify a fuzzy

number [14], which is achieved by mathematical integral. Therefore, P 's totally satisfactory degree of keywords is 0.8508.

The search results are ordering based on several fuzzy factors including the totally satisfactory degrees of keywords, the page ranks, the page in-links, titles, and last modified dates. Through the mutual compensation of all fuzzy factors, an ordering of search results which is considered in every respect can be generated. The fuzzy weighted average [8,11] is utilized to aggregate all fuzzy factors based on their degrees of importance and degrees of satisfaction. The users can manually select appropriate fuzzy factors with different importance degrees to meet their specific requirements (see Figure 11) -- otherwise the importance degrees of fuzzy factors would be default values. Figure 12 shows the ordering of search results which the grade is generated by aggregating all fuzzy factors.



Figure 11. The selection of fuzzy factors with different importance degrees

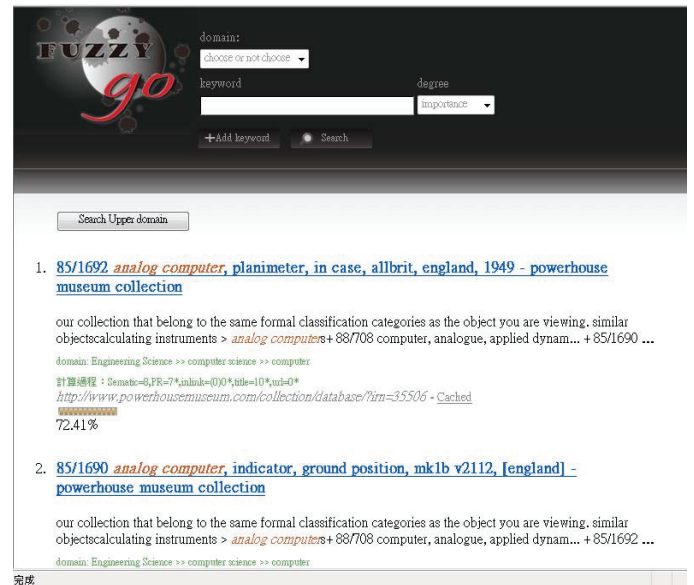


Figure 12. The ordering of search results

However, the original setting is not optimal, and it needs continuous learning and adaptation to fit users' expectations gradually. We have applied the genetic algorithm to propose a self-adaptation approach to Fuzzy-Go search engine [12]. For each search, the fuzzy search engine records the difference between the ordering of search results and user's real behavior on clicking web pages. This feedback is gathered and analyzed to adjust the fuzzy similarities between terms in the fuzzy ontology, the domain classification of web pages, and the default importance degrees of fuzzy factors. The ordering of search results can thus be improved gradually by continuous learning and adaptation.

V. CONCLUSION

In this paper, we develop a fuzzy search engine, called Fuzzy-Go. First, the Ontology Manager constructs and maintains a Fuzzy Ontology. A data mining approach is applied on the Web Document to calculate the fuzzy similarity between terms in the Fuzzy Ontology. The Fuzzy Ontology is constructed by using fuzzy logic to capture the similarities of terms in the ontology, which offering appropriate semantic distances between terms to accomplish the semantic search of keywords. The Fuzzy-Go search engine can thus automatically retrieve web pages that contain synonyms or terms similar to keywords. Second, the Web Crawler is developed to gather and classify web pages. Web pages are classified and stored based on their domains. To deal with the huge amount of retrieved web pages, we propose a data mining approach to cluster web pages based on their keywords. Third, the User Interface provides users to select an appropriate domain and to input multiple keywords with different degrees of importance based on their needs. The totally satisfactory degree of keywords can be aggregated based on their degrees of importance and degrees of satisfaction. Finally, the Fuzzy Search Mechanism would exclude web pages in the inappropriate domains to reduce the search space and to improve the search results. The keywords are expanded by the Fuzzy Ontology to find out synonyms and terms similar to keywords. The search results are ordering based on several fuzzy factors including the totally satisfactory degrees of keywords, the page ranks, the page in-links, titles, and last modified dates. The advantages of the proposed approach are as follows.

- Using fuzzy logic to capture the similarities of terms in the ontology, which offers appropriate semantic distances between terms to accomplish the semantic search of keywords. The domain classification of web pages offers users the selection of the appropriate domain to search web pages, which excludes web

pages in the inappropriate domains to reduce the search space and to improve the search results.

- Users can input multiple keywords with different degrees of importance based on their needs, and all fuzzy factors are aggregated based on their degrees of importance and degrees of satisfaction. Through the mutual compensation of all fuzzy factors, the ordering of search results which is considered in every respect can be generated.

ACKNOWLEDGMENT

This research is partially sponsored by National Science Council (Taiwan, R.O.C.) under the grant NSC 98-2221-E-018-008-MY2.

REFERENCES

- [1] Keyword Density, http://en.wikipedia.org/wiki/Keyword_density
- [2] Page Inlink Analysis, <http://ericmiraglia.com/inlink/>
- [3] Page Rank Check, http://www.prchecker.info/check_page_rank.php
- [4] The Stochastic Language Models (N-Gram) Specification, <http://www.w3.org/tr/ngram-spec/>
- [5] The Suggested Upper Merged Ontology (SUMO), <http://www.ontologyportal.org/>
- [6] WordNet, <http://wordnet.princeton.edu/>
- [7] J. C. Bezdek, Pattern Recognition with Fuzzy Objective Function Algorithms. New York: Plenum, 1981.
- [8] P.T. Chang, K.C. Hung, K.P. Lin, and C.H. Chang, a Comparison of Discrete Algorithms for Fuzzy Weighted Average, *IEEE Transactions on Fuzzy Systems*, pp.:663-675, Oct. 2006.
- [9] K.W. Church and P. Hanks Word Association Norms, *Mutual Information and Lexicography*, *Computational Linguistics* 16(1):22-29, Mar. 1990.
- [10] D. Dubois and H. Prade. *Fuzzy sets and systems: theory and applications*. New York, London, 1980.
- [11] L.F. Lai, C.C. Wu, M.Y. Shih, L.T. Huang, and W. Chiou. Parallel Processing for Fuzzy Queries in Human Resources Websites. *Journal of Internet Technology*, 7(11):943-953, Dec. 2010.
- [12] Y.C. Lin, L.F. Lai, C.C. Wu, and L.T. Huang. A Self-Adaptation Approach to Fuzzy-Go Search Engine. *The 2010 International Computer Symposium (ICS 2010)*, pp. 1020-1025, Dec. 2010.
- [13] E.W.T. Ngai and F.K.T. Wat. Fuzzy Decision Support System for Risk Analysis in E-Commerce Development. *Decision Support Systems*. pp.:235-255, Aug. 2005.
- [14] T.Y. Tseng and C.M. Klein. A New Algorithm for Fuzzy Multicriteria Decision Making. *International Journal of Approximate Reasoning*, 6:45-66, 1992.
- [15] L.A. Zadeh. Fuzzy Sets. *Information and Control*, pp.338-353, 1965.
- [16] H.J. Zimmermann. *Fuzzy set theory and its applications*. 2nd revised edition, Kluwer Academic Publishers, 1991.