

Fuzzy Semantic Search Engine

Dharmish Shah

Dept. of Computer Engineering
Dwarkadas J. Sanghvi College of
Engineering
Mumbai, India

Jheel Somaiya

Dept. of Computer Engineering
Dwarkadas J. Sanghvi College of
Engineering
Mumbai, India

Sindhu Nair

Assistant Professor
Dept. of Computer Engineering
Dwarkadas J. Sanghvi College of
Engineering
Mumbai, India

ABSTRACT

Almost all the search engines that exist retrieve web pages by finding the exact keywords. The traditional keyword-based search engines suffer many problems, like synonyms and terms similar to keywords are not taken into account to search web pages, they treat all keywords as the same importance and cannot differentiate the importance of one keyword from that of another.

Synonyms and terms similar to keywords are not taken into consideration to search web pages. Users may need to think of and 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 that the users face, in this paper we have proposed and applied the fuzzy logic theory and the semantic search techniques to develop a fuzzy semantic search engine.

General Terms

Web Search, Search Engine, Query Evaluator.

Keywords

Fuzzy Search Engine, Fuzzy Ontology, Semantic Search.

1. INTRODUCTION

Nowadays the volume of the information on the Web is increasing drastically. Facilitating users to get useful information has become most important to information retrieval systems. While information retrieval technologies have been improved to some extent. With the emergence of the Semantic Web, this situation can be remarkably improved if machines could “understand” the content of web pages. The existing information retrieval technologies can be classified mainly into three classes. The traditional information retrieval technologies are almost based merely on the occurrence of words in documents. It is only limited to string matching. However, these technologies are of no use when a search is based on the meaning of words, rather than on words themselves.

The most widely used algorithms are the PageRank algorithm and the HITS algorithm. The widespread availability of machine understandable information on the Semantic Web offers some opportunities to improve traditional search. If machines could “understand” the content of web pages, searches with high precision and recall would be possible.

Fuzzy Semantic Search Engine falls in the area of information retrieval and semantic web, and aims to improve the evaluation and enhances web search tools. The huge number of information on the web as well as the growth of new inexperienced users creates new challenges for information retrieval; certainly the current search engines offer an efficient way to browse the web content. However, these types of tools do not take into account the semantics driven by the query terms and document words. This paper proposes a new semantic based approach for the evaluation of information retrieval systems; the goal is to increase the selectivity of search tools and to improve how these tools are evaluated. The user will have the option to search for the information as per the user needs. The results will be shown not only based on the keywords but also on the related pages relevant to the user. The users will get only the relevant information for what they are searching. The activities for searching the information on the web by the user will be recorded for the maintenance and analysis purpose. Based on this recorded information the user search related pages will be shown based on the time they have stayed on the visited pages previously.

2. CONCEPTUAL THEORY

Semantic Web search intends to improve the accuracy of search by understanding search intent and the contextual meaning of terms as they appear in the searched data space to generate more accurate and relevant results. Semantic search systems consider various ideas including contextual search, location, synonyms, intent, and variation of the words, generalized and specialized queries, conceptual matching and natural language queries to provide relevant search results. Major web search engines namely Bing and Google incorporate some elements of semantic search. Rather than using various ranking algorithms such as Google's PageRank to predict the relevancy, semantic search uses semantics or the science of meaning in language, to produce greatly relevant search results. Commonly used searching methodologies for semantic search are Logics - by using inference based on OWL, Fuzzy concepts, fuzzy relations and fuzzy logics. Ontology can be defined as a systematic description of part-of relationships and entity dependencies. In other words, ontology consists of a hierarchical description of important classes (or concepts) in a specified domain, along with the description of the properties (of instances) of each and every concept. The Web Ontology Language (OWL) is a family of knowledge representation languages for authoring ontologies. Description Logics (DL) are a family of knowledge representation languages which can be used to represent the terminological knowledge of an application domain in a structured and well-understood formal way. Today description

logic has become a cornerstone of the Semantic Web for its use in the design of ontologies. There is no unique definition of fuzzy ontology. In the simplest case, a Fuzzy Ontology is a pair (C, R), where C is a set of (fuzzy) concepts and R is a set of (fuzzy) binary (n-ary) relations. Fuzzy ontology can be seen as an extended domain ontology which makes use of the specific domain and fuzzy information processing as follows: (i) the input is unstructured data; (ii) the definition in the domain of related concepts, e.g., objects, instances, and their relationships; (iii) the generation of domain ontology; (iv) the domain ontology extended as fuzzy ontology; and (v) applying the fuzzy ontology to the specific domain.

3. PROPOSED SYSTEM

We propose a Search Engine that understands the meaning of the user query and relatively reasons him with the appropriate result. (See figure 1) Not only the user entered keyword based pages would be returned but also the pages that is appropriate enough with the meaning of the user entered keyword. We will also provide user with facility to mark pages which would be displayed first the next time user enters the terms that are related. The User Interface of the fuzzy search engine provides users to select an appropriate domain and to input multiple keywords with different degrees of importance based on their needs. On the one hand, web pages in the inappropriate domains can be excluded to reduce the search space and to improve the search results, since all web pages are already classified by domains. On the other hand, a better search result can be obtained by differentiating between keywords according to the degrees of importance.

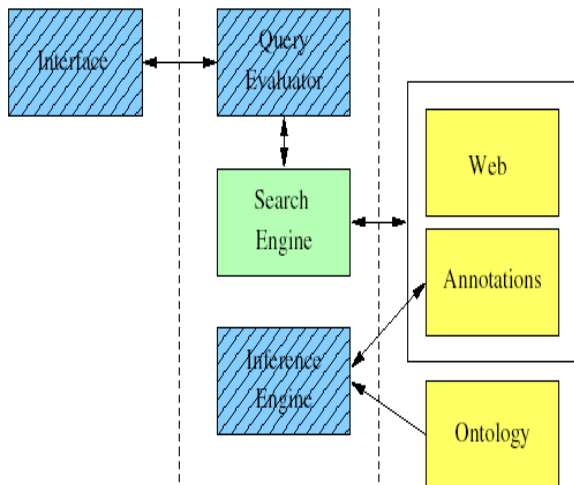


Fig 1: Proposed System

Interface: The user interface through which user will enter phrases or keywords to search.

Query Evaluator: It block filters out the keywords from the user entered phrases and generates the synonyms to it.

Search Engine: It takes the keywords from the query evaluator and checks it in the web document for the relevant pages which is returned to the inference system.

Inference System: Using background ontology Inference system adds all properties that can be deduced / induced from the ontology and returned to the web documents for other relevant pages.

Time Rank Mechanism: We try to implement a Time Rank Mechanism for ranking the pages which user searches. This is a simple mechanism which ranks the pages based on the amount of time user has stayed on it previously. Higher the time, higher would be the rank of the page.

We develop a crawler agent to automatically extract web pages and inventory them by finding out keywords and characteristic values. 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) 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

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

4. CONCLUSION

In this paper, we develop a fuzzy search engine, called Fuzzy Semantic Search Engine. Firstly, 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 offer appropriate semantic distances between terms to accomplish the semantic search of keywords. The Fuzzy Semantic search engine can thus automatically retrieve web pages that contain synonyms or terms similar to keywords. Secondly, the Web Crawler is developed for gathering and classifying web pages. Web pages are stored and classified based on their specific domains. To deal with the huge amount of retrieved web pages, we propose a data mining approach to cluster web pages based on the keywords entered by the user.

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 total satisfactory degree of keywords can be aggregated based on their degrees of satisfaction and degrees of importance. 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 total satisfactory degrees of keywords, the page in-links, the page ranks titles, and last modified dates.

5. ACKNOWLEDGEMENT

The authors wish to acknowledge the Department of Computer Engineering, Dwarkadas J. Sanghvi College of Engineering for their support and technical help throughout this work.

6. REFERENCES

- [1] Lien-Fu Lai, Chao-Chin Wu, Pei-Ying Lin, “Developing a Fuzzy Search Engine Based on Fuzzy Ontology and Semantic Search”. Dept. of Computer Science and Information Engineering National Changhua University of Education Changhua, R.O.C.
- [2] http://en.wikipedia.org/wiki/Web_Ontology_Language
- [3] http://en.wikipedia.org/wiki/Semantic_search
- [4] <http://gaia.isti.cnr.it/straccia./software/FuzzyOWL/index.html>
- [5] J. C. Bezdek, Pattern Recognition with Fuzzy Objective Function Algorithms. New York: Plenum, 1981.
- [6] 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.
- [7] K.W. Church and P. Hanks Word Association Norms, Mutual Information and Lexicography, Computational Linguistics 16(1):22-29, Mar. 1990.
- [8] D. Dubois and H. Prade. Fuzzy sets and systems: theory and applications. New York, London, 1980.
- [9] 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.
- [10] 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.
- [11] 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.