ISSN (Online) : 2278-1021
ISSN (Print) : 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 3, Issue 4, April 2014

Yet another way of Ranking web Documents Based On Semantic Similarity

Vidya Kannan[1], Dr. G.N Srinivasan[2]

Department of Information Science and Engineering, RV College of Engineering, Bangalore, India[1,2]

Abstract: In today's web enabled world, searching for relevant information on web has been an important topic of research. Semantic similarity aims at providing robust tools for standardizing the content and delivery of relevant information across communicating information sources. Most of the times the user gets lots of irrelevant data as a result of poorly implemented search process. To avoid this, a ranking scheme is proposed, which provides the search result-set according to the better understood and correctly interpreted user query. This is done by considering the relevance of the query by keeping the user view in mind and also the semantics of the document and the user query. The simple lexical and/or syntactical matching usually used by search engines does not extract web documents to the user expectations. The proposed solution provides the most relevant data to user ranked in their relevance. The proposed ranking scheme for the semantic web search engine functions by finding the semantic similarity between the information available on the web and the query which is specified by the user. This approach considers both the syntactic structure of the document and the semantic structure of the document and the query. The objective of this paper is to demonstrate that a semantic similarity based ranking scheme will provide much better results than those by the prevailing methods. In this Paper an algorithm will be implemented that provides ranking scheme for the semantic web documents by finding the semantic similarity between the documents and the query which is specified by the user. The algorithm considers both syntactical and semantic similarities of the query and categorizes the search results based on the most probable and most appropriate interpretation of the query based on various interpretations taking into account all the words and their combinations in the query.

Keywords: semantic similarity, Ontology, IDFT

I. INTRODUCTION

Semantic similarity measure plays an important role in semantic search. In this paper, we proposed ontology based semantic similarity method to rank the web documents. The process of web document classification involves calculating similarities between documents and categories by using the information extracted from them. In recent years, ontology-based web documents classification method is introduced to solve the problem of classifier training and not considering semantic relations between words in traditional Machine Learning algorithms. However, previous works on ontology-based web documents classification miss some important issues of automatic ontology construction and ranking of classified documents. In order to solve these problems, this paper proposes an ontology-based web documents classification and ranking method. Firstly, weighted terms set are extracted from web documents, and ontology is build up by using an effective ontology construction method which clarifies and augments an existent ontology. In this paper a ranking scheme is proposed for the semantic web documents by finding the semantic similarity between the documents and the query which is specified by the user. The novel approach proposed here not only relies on the syntactic structure of the document but also considers the semantic structure of the document and the query. The combined use of conceptual, linguistic and ontology based matching has significantly improved the performance of the proposed ranking scheme.

II. RELATED WORK

In fact, the information retrieval process by a search engine is very crucial. There are many ranking models [7] that have been proposed by the various researchers like Boolean model [6], statistical model [9], Hyperlink based model [3], Conceptual model [16] and many more [4, 8] which has been widely used. Some of them use the natural language processing techniques such as language model and relaxation algorithm. The use of natural language techniques in these models helps to consider syntactic, semantic structure and morphological form of terms. Some other document ranking models such as Neural Networks, Fuzzy Sets, Relevance Feedback Models could be used for efficiently increasing the performance of ranking models.There are two kinds of methods to measure the semantic similarity of two keywords in ontology:

Elemental based matching and Structure-based matching [3]. AgentMatcher [4] is a decision-making system in elearning environment ,which uses the matching algorithm to match the service provider and user query.The matching algorithm was deployed in edusource, the famous education project in Canada learning object metadata (CanLOM), as a learning object searching component. Agent matcher consists of the user interface, translation tools, metadata Generator, similar measurements engine. Similarity measurement engine is the core module of the AgentMatcher, the main idea is to computing the similarity of weighted trees based on the term taxonomy. It use Elemental based matching method to solve the xml based tree matching problem.

In order to solve these problems, some ontology-based classification methods are researched [3, 4 and 5]. These methods use ontologies to represent characteristics of the categories, so web documents are classified in real time

ISSN (Online) : 2278-1021
ISSN (Print) : 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 3, Issue 4, April 2014

not with training data or a learning process, as ontologies become more detailed by evolution process, the classification method can achieve better precision and recall; when calculating the similarity score between web documents and categories, ontology-based classification method consider the semantic relations between the terminology information extracted from web texts and ontology categories.

III. PROPOSED RANKING MODEL

In this project a ranking scheme is proposed for the semantic web documents by finding the semantic similarity between the documents and the query which is specified by the user. The novel approach proposed here not only relies on the syntactic structure of the document but also considers the semantic structure of the document and the query. The combined use of conceptual, linguistic and ontology based matching has significantly improved the performance of the proposed ranking scheme. A ranking scheme is proposed for the semantic web documents by finding the semantic similarity between the documents and the query which is specified by the user. This approach considers both the syntactic structure of the document and the semantic structure of the document and the query. The approach used here includes the lexical as well as the conceptual matching. The combined use of conceptual, linguistic and ontology based matching has significantly improved the performance of the proposed ranking scheme. The semantic similarity based ranking scheme gives much better results than those by the prevailing methods is been found here.

In this Project an algorithm will be implemented that provides ranking scheme for the semantic web documents by finding the semantic similarity between the documents and the query which is specified by the user. In the Existing System the website is searched based on the word present in the documents. i.e a search engine would consider even unnecessary words known as stop words. The system does not even have capability in order to customize search in which user can add his own words to remove during the customization process. In the previous approaches the Website selection was done manually by a review software and the set of relevant website are chosen. In this project phase1 and phase2 are implemented in which the Websites are mined and converted into tokens. The frequencies of all tokens are found out for each of the Website approvals and finally text document encoding is done by using Inverse Document Frequency algorithm. Finally all Website proposals are listed based on the rank.

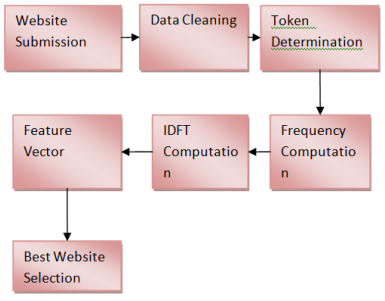

Figure 1: Architecture and system flow of proposed ranking model

Algorithm:

Website Submission: This module is responsible to validate the websites and then access the data from the website and remove the HTML DOM elements and produce a data without HTML content.

Data Cleaning- This module is used in order to remove stop words from the website.

Tokenization's- This process in used to obtain all the keywords of the website and assign them a unique ID as well as the web site id.

Text Frequency – This module is used to compute the text frequency i.e no of times a token appears in the Website.

IDFT- This is used to compute IDFT by taking the ratio of Text Frequency to the Number of Documents.

Feature Vector: This is multiplication of Text Frequency and IDFT Frequency.

Ranking Website- This is used to rank the Websites based on the Ascending order of the Feature Vector value.

Feature vector= Fi*Idft

Idft= log(N/fi)

Where Fi=token frequency

N= no: of web pages in which the particular token appears

IV. PERFORMANCE ANALYSIS

Performance of our approach for finding semantic similarity between the semantic web documents and the query considers not only on the keywords but also on the associated concept.

Relations that exists between the concepts that are extracted from the document. This gives more specific similarity of the document with the user query. Here the frequency of each token is calculated. Then the corresponding Idft value is calculated.

The diagram below shows the graphical representation of frequency versus feature vector id and idft versus feature vector id.

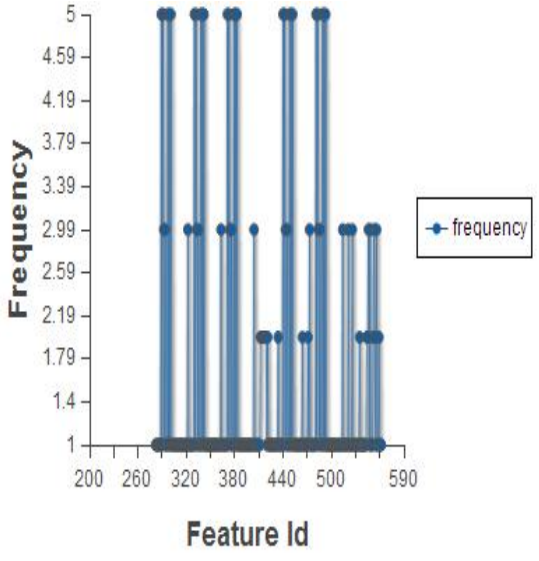

Figure 2

ISSN (Online) : 2278-1021
ISSN (Print) : 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 3, Issue 4, April 2014

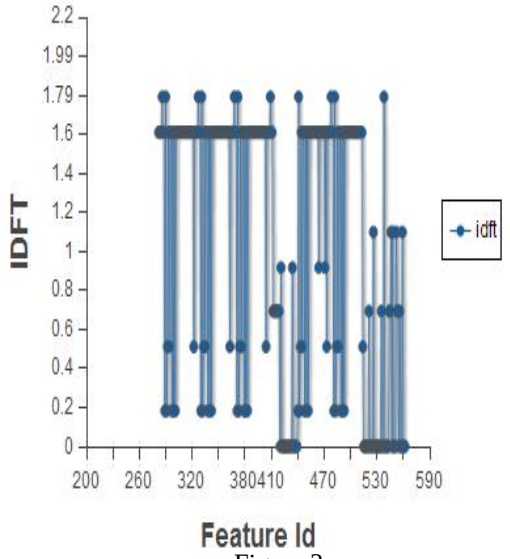

Figure 3

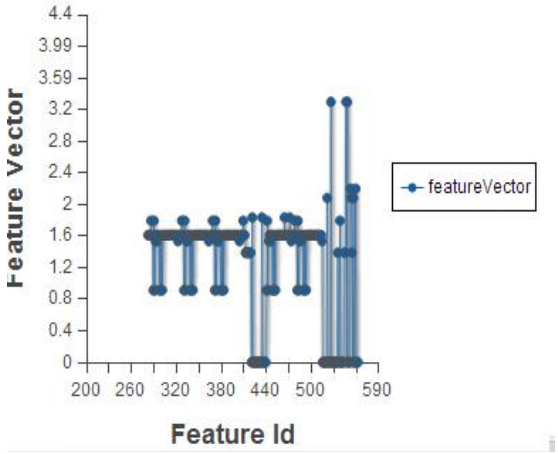

Figure 4

V. CONCLUSION AND FUTURESCOPE

The above mentioned ranking model deals with improving the search strategies several ways , thus retrieves the most relevant pages. This ranking model takes the concepts and relationship between the concepts which exists both in the document and user query to improve the retrieval of relevant document. Our future efforts would be to design more meaningful and exhaustive ranking strategy by using the semantic analysis of web pages and by deeply statistical analysis relevance of documents, so that the semantic search engine can evaluate more precisely relevance and also the similarity between the web page and the user query. The ranking can even be done by any ontology already created or automatically creating a new ontology for the documents and the user query and then comparing them for the relevance score. We will also try to make our approach scalable for the semantic web.

REFERENCES

[1] Berners-Lee T., Hendler J., and O. Lassila, "The Semantic Web," Scientific Am., 2001.

[2] Bollegala D., Matsuo Y., and Mitsuru, "A Relational Model of Semantic Similarity between Words using Automatically Extracted Lexical Pattern Clusters from the Web', Proc of Int'l Conf on Empirical Methods in Natural Language Pr ocessing,pp 803-812, August 2009.

[3] Bollegala D., Matsuo Y., and Mitsuru, "A Web Search Engine-Based approach to measure Semantic Similarity between Words", IEEE Transactions on knowledge and Data Engineering, vol. 23, no. 7, pp 977-990,July 2011.

[4] Cosma G. and Mike, "An approach to source-code Plagiarism Detection and Investigation using Latent Semantic Analysis",IEEE Transaction on Computers, vol. 61, no. 3, pp 379-394, March 2012.

[5] Ding L., Kolari P., Ding Z., and S. Avancha, "Using Ontologies in the Semantic Web: A Survey", Ontologies,integrated series of information systems, vol 14, pp. 79-113,Springer, 2007.

[6] E. Greengrass, "Information Retrieval: A survey". DOD Technical

[7] Report TR-R52-008-001, November 2000.

[8] Grossman D., and O. Frieder. "Information retrieval algorithms and heuristics". Second ed. . Springer. 2004.

[9] Iosif E., and Potamianous, "Unsupervised Semantic Similarity Computation Between Terms usingWeb Documents", IEEE Transaction on Knowledge and Data Engineering,vol. 22, no. 11, pp 1637-1647, November 2010.

[10] Lempel R., S. Moran. "The stochastic approach for linkstructure analysis (SALSA) and the TKC e®ect". In The Ninth International WWW Conference, May 2000.

[11] Li Z., and Karthik R., "Ontology-based Design Information Extraction and Retrieval", Artificial Intelligence for Engineering Design, Analysis and Manufacturing, vol 21, pp 137-154, 2007.

[12] Oleshchuk V., and Asle P., "Ontology Based Semantic Similarity Comparison of Documents", Proc. of IEEE 14th workshop on database and expert systems applications,2003.

[13] Page L., S. Brin, R. Motwani, and T. Winograd, "The Page Rank Citation Ranking: Bringing Order to the Web", Stanford Digital Library Technologies Project, 1998.

[14] Protiti M., "Semantic web: The future of WWW", Proc. Of 5th Int'l Conf. CALIBER, Punjab University, Chandigarh, 08-10,2007.

[15] Shamsfard M., Namehtzadeh A., and S. Motiee "ORank: An Ontology based System for Ranking Documents", Int'l Journal of Computer Science, vol 1, no 3, ISSN 1306-4428, 2006.