TERM BASED WEIGHT MEASURE FOR INFORMATION FILTERING IN SEARCH ENGINES

Mu. Annalakshmi
Research Scholar, Department of Computer Science, Alagappa University, Karaikudi.
annalakshmi_mu@yahoo.co.in

Dr. A. Padmapriya
Associate Professor, Department of Computer Science, Alagappa University, Karaikudi.
mailtopadhu@yahoo.co.in

Abstract - The World Wide Web has been loaded with enormous amount of data in the recent years. Web Mining is the application of data mining techniques to find interesting and potentially useful knowledge from web data and hence it has gained a significant importance. Information filtering (IF) has recently emerged as a technique for effective delivery of the required and also relevant information. Most of the information filtering in web is being done with the help of search engines which provide the user with large number of documents matching the search query. The users have to filter out the relevant documents from the search results, which is a cumbersome process. Various methods have been developed for information filtering in search engines using ontology. The proposed system retrieves the results from the search engine and ranks them in the decreasing order of relevance. The relevance is determined by the occurrence of the search query terms or their corresponding synonyms in the HTML source code of the webpage. Different weights are assigned to the matching terms in URL of the web page and different tags such as title, meta tags, headings and image tags. Weights are also given to the occurrence of the keywords/synonyms in the body of the code as free text and also their nearness. A relevance score is calculated for each of the web pages based on the weights which determine its position in the search results. The experimental results are analyzed with the popular search engine Google. The proposed method filters more relevant pages of the search results at the top of the list thereby helping the users in finding their information need in a more easy way.

Keywords : Information Filtering, Web mining, Search engine, Relevance score

1. INTRODUCTION

The World Wide Web has been loaded with enormous amount of data in the recent years. So finding useful information from such large data repository has become more and more difficult with such huge increase in data.

Web Mining is the application of data mining techniques to find interesting and potentially useful knowledge from web data and hence it has gained a significant importance. Mining web has been difficult due to the following reasons.

1. Vast size of the web
2. Diverse nature of the web since it consists of text, images, video, audio and other multimedia content.
3. Exponential increase in size of web due to addition of new information.
4. Dynamic nature of web since the most of the web content is modified frequently.
5. Duplicate content in the web.
6. Hyperlinks to other web documents.
7. Unstructured or semi-structured nature of information.

Web Mining is classified into three types namely

- Web Content Mining
- Web Usage Mining
- Web Structure Mining

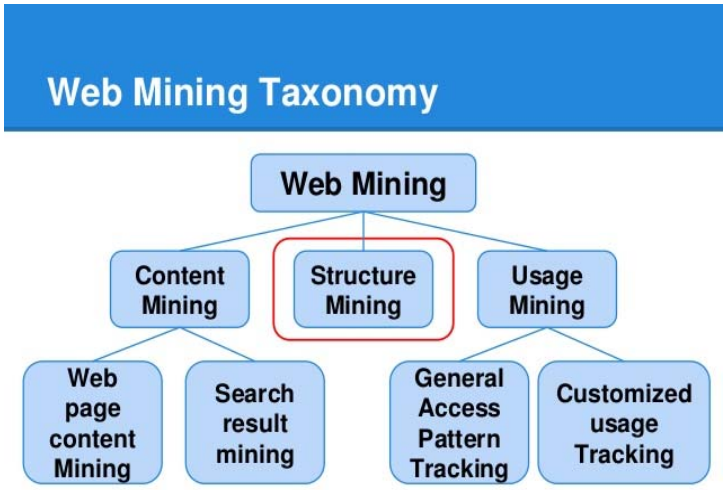

Fig 1. Taxonomy of Web Mining

Web Content Mining:

Web content mining [2] extracts or mines useful information or knowledge from web page contents. Web content is a very rich information resource consisting of many types of information such as free text, image, audio, video, metadata as well as hyperlinks. Web pages are mostly semi-structured (e.g., HTML documents) or unstructured free text data. Semi- structured documents use the HTML structure for mining data. On the other hand, unstructured free text use bag of words approach, which is based on the occurrence of words.

Web Usage Mining:

Web usage mining [2] refers to the discovery of user access patterns from web usage logs, which record every click made by each user. The mined data usually deals with web server logs, proxy server logs and browser logs. These logs include information about referring pages, user identification, time spent by the user at a site and the order in which the pages are visited. This information helps in reorganization of the website to better suit the user.

Web Structure Mining:

Web structure mining [2] discovers useful knowledge from hyperlinks which represent the structure of the Web. Graph theory is basically used by web structure mining for analysis of connection structure of web site. This helps in discovering important sites for a particular topic or discipline or in discovering environments.

The paper is organized as follows. The background study is elaborated in section 2. The related works which act as motivating factor for the proposed work are described in section 3. The proposed method is explained in section 4. The experimental study is summarized in section 5. Section 6 concludes the research work.

2. BACKGROUND STUDY

Information filtering (IF) has recently emerged as a technique for effective delivery of the required and also relevant information. Some of the features of IF systems are

- can be applied to unstructured or semi-structured data such as documents, e-mails, etc.
- can handle large amounts of data.
- deal primarily with textual data
- are based on user profiles
- remove irrelevant data from incoming streams of data.

A. SEARCH ENGINES

Most of the information filtering in web is being done with the help of search engines. Search engine[2] is a tool that helps the web users to search the required information from the voluminous data available in the web. In other words, it is a program that searches documents for specified keywords and return a list of matching documents. In [3], the author discusses about various popular search engines such as Google, Bing, yahoo and their ranking factors. The users search the web using queries which are usually short and a large number of web

pages are returned to them as the search result. The users have to go through these web pages and find the required information which is difficult and time-consuming.

B. SEMANTIC WEB

Tim Berners-Lee, the inventor of the World Wide Web described the Semantic Web as an extension of the current web in which information is given well-defined meaning, better enabling computers and people to work in cooperation.

Semantic web mining combines two popular research areas, semantic web and web mining. It aims to improve the quality of web mining by making use of the semantic structure of the web pages. Ontology forms the backbone of semantic web. (The core of semantic web is ontology). It refers to the description of concepts in a domain as classes along with their properties and relationships.

WordNet is a lexical database in English which mainly stores the synonyms of a word (when used as noun, verb, adverb or adjective) along with its related words such as hypernyms, hyponyms, holonyms, meronyms and their corresponding meanings. This entire set is the called the synset of a particular word.

Semantic Web Mining and Ontology in particular can be used to filter the user's required information easily by considering the keywords, their synonyms and words related to them.

3. RELATED WORK

Various methods have been developed for information filtering in search engines using ontology.

Ontology content-based filtering Method[4] measures the similarity between user profiles and the content representation of documents and also orders the documents according to the relevancy to each user.

Yan Shen et al. [5] proposed an ontology based approach to construct levels for matching user search intents. It uses iterative mining algorithm for evaluating potential intents level by level until getting the best result.

The Personalized Web Search by User Learned User Profiles in Re-ranking[6] provides an improved scoring function by using term characterization, image characteristics and pivoted length normalization for re-ranking of search results.

Kwang Mong Sim[8] proposed an Information Filtering Agent(IFA) which determines the relevancy of web pages by assigning weights to evidence phrases, frequency of evidence phrases and their nearness. Evidence phrases refer to exact keywords, their synonyms, hypernyms and hyponyms. The documents are searched for the occurrence of evidence phrases after the removal of function words. The relevance of the web pages is determined by these three heuristics.

4. PROPOSED SYSTEM

The proposed system retrieves the results from the search engine and ranks them in the decreasing order of relevance. The relevance is determined by the occurrence of the search query terms or their corresponding synonyms in the HTML source code of the web page. Different weights are assigned to the matching terms in URL of the web page and different tags such as Title, Meta tags, headings, and image tags .Weights are also given to the occurrence of the keywords in the body of the code as free text and also to the nearness between the keywords. A relevance score is calculated for each of the web pages based on the weights which determines the order in the results. The architecture of the proposed system is shown below.

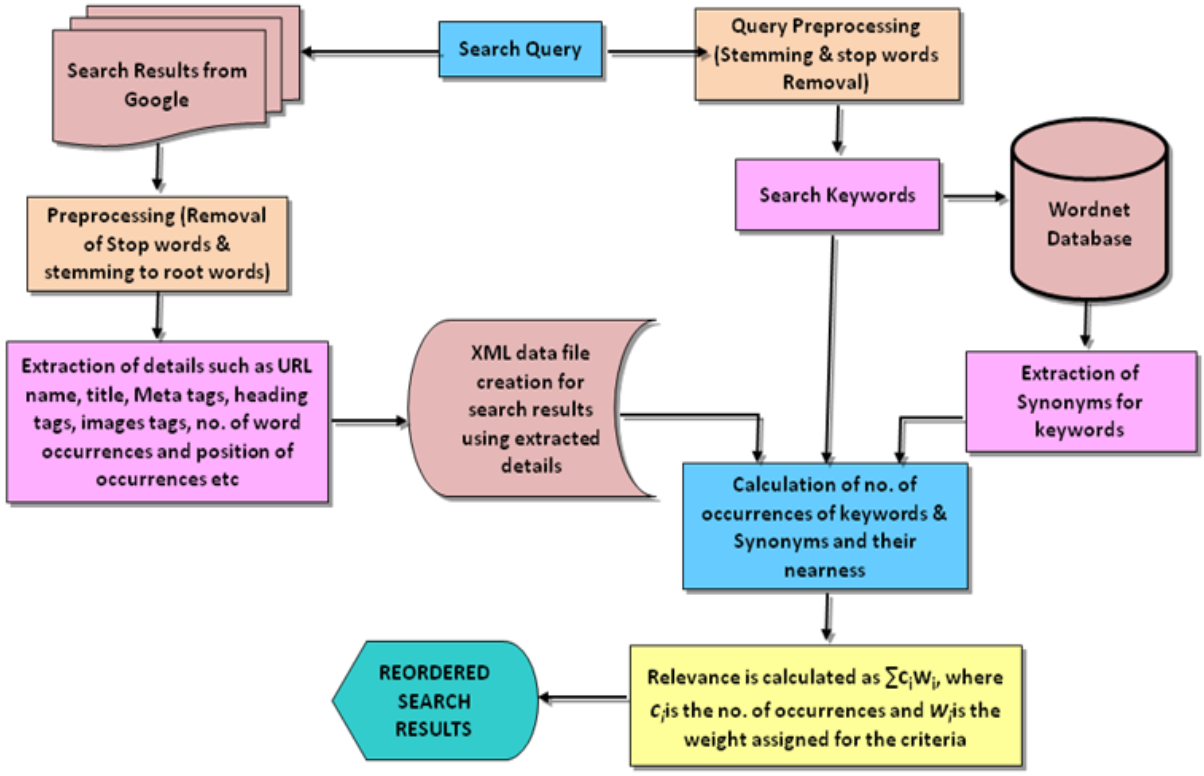

Fig 2. Architecture of the proposed system

Algorithmic steps:

1) The search query is given to the search engine and the results are retrieved from the top 30 websites of Google.
2) An XML file is created for each of the retrieved web pages. The details of the web pages such as URL name, title, meta tags, heading tags, image tags and number of word occurrences and their position of occurrence for each of the words in the webpage are retrieved after pre-processing which involves the removal of stop words and stemming to root words.
3) The words in the search query are also pre-processed.
4) The synonyms for each of the keywords in the search query are retrieved from Wordnet and pre-processed.
5) The count of occurrence each of the keywords in the query and also their synonyms in URL, Title, Meta tags, heading1, image tag and normal text are calculated. In case of normal text their nearness is also taken into account.
6) Weights are assigned to each of these occurrences.
7) The relevance of the web page to the given query is calculated by sum of product of the count and weight of the keywords and their respective synonyms.
8) Based on the relevance the search results are displayed in the decreasing order of relevance.

5. EXPERIMENTAL STUDY

The proposed method was evaluated by collecting sample data from 150 students belonging to under graduate, post-graduate, master of philosophy and doctoral scholars in the field of computer science. They were given five queries and they were asked to evaluate the relevance of top 30 search results of Google for each query. Based on the average of the relevance feedback obtained from the students, it was found that the proposed ranking method ranked more relevant pages at the top 15 results when compared with Google. The number of relevant pages in the last 5 search results is another parameter which indicates how deep the user has to browse the search results. This number was less in the proposed method.

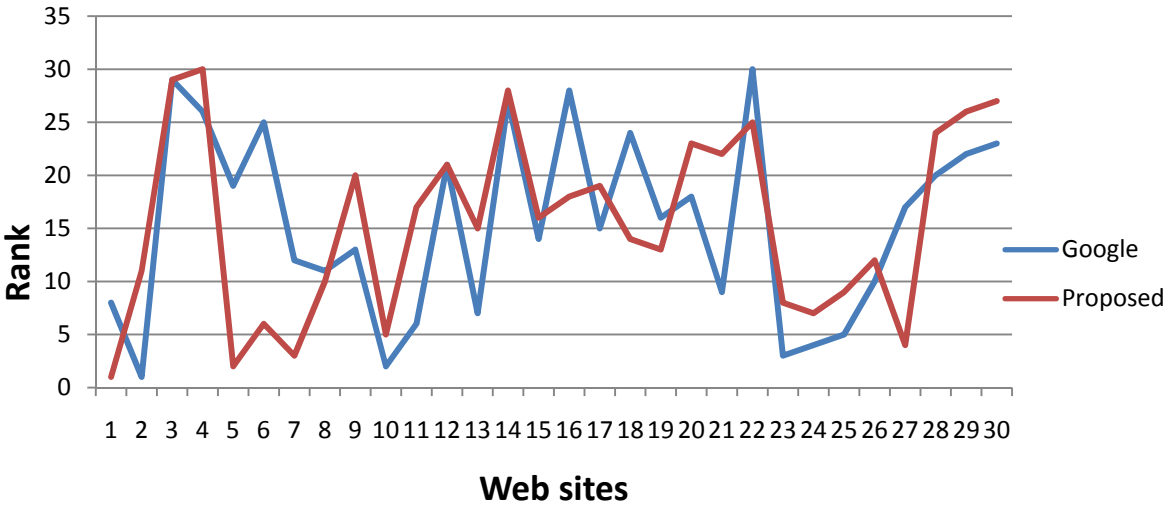

Fig 3. Ranking of websites by Google and Proposed Method for the search query “Pointers in C”

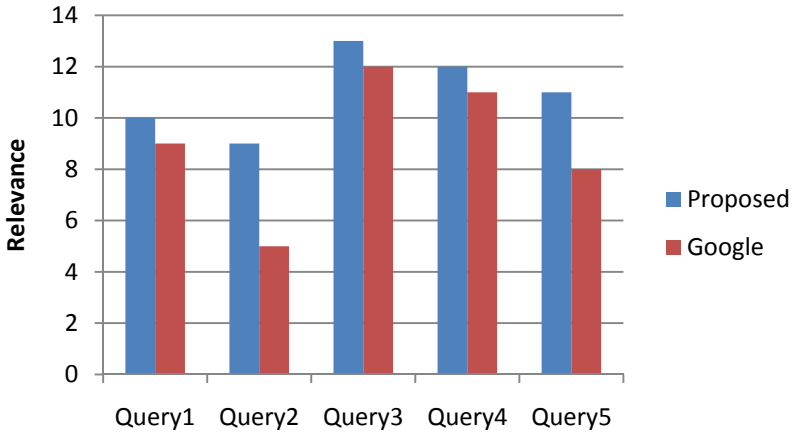

Fig. 4. Number of Relevant pages in Top 15 Results

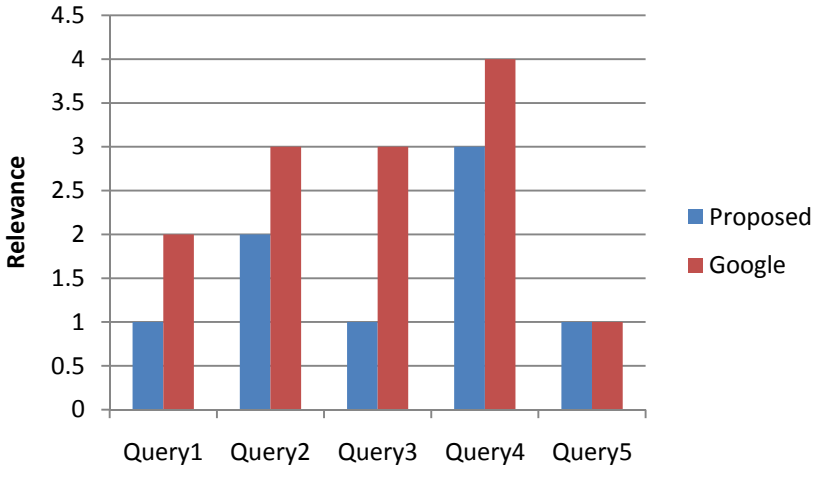

Fig. 5. Number of Relevant pages in Last 5 Results

6. CONCLUSION

This paper proposes a term based weight measure to the user for retrieving the required information from a large number of websites listed out by the search engines. The weights assigned to the occurrence of the word or synonym in different places of the HTML source code determines the ranking. This method is found to be closer to the user ratings than that of Google. As of now, the sample queries used for evaluation were restricted to the field of computer science. In future, the general search queries may be used for evaluation of the proposed method.

REFERENCES

[1] https://en.wikipedia.org/wiki/Web_mining: Last accessed on 26th February 2017
[2] G.K. Gupta, "Introduction to Data Mining with case studies", Prentice Hall of India
[3] Mu. Annalakshmi, Dr.A. Padmapriya, "Search factors used by search engines", International Journal of Advanced Research Trends in Engineering and Technology", Vol. 3, Special Issue 20, April 2016.
[4] Peretz Shoval, Veronica Maidel, Bracha Shapira, "An Ontology content-based filtering Method, International Journal on Information Theories & Applications", Vol.15/2008.
[5] Yan Shen, Yuefeng Li, Yue Xu, "An Ontology-based Mining Approach for User Search Intent Discovery", Proceedings of the 16th Australian Computing Symposium, Canberra, Australia, 2 Dec 2011.
[6] Jia Hu and Philip K.Chan, "Personalized Web Search by User Learned User Profiles in Re-ranking" , Workshop on Knowledge Discovery on the Web, KDD conf. pp. 84-97, 2008
[7] Valerio Basile, "WordNet as an Ontology for Generation" WebNLG 2015 1st International Workshop on Natural Language Generation from the Semantic Web, Jum 2015, Nancy, France.
[8] Kwang Mong Sim, "Toward an Ontology-Enhanced Information Filtering Agent" in ACM SIGMOD Rec., Vol. 33, Mar 2004.