International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 07 Issue: 02 | Feb 2020 www.irjet.net p-ISSN: 2395-0072

Re-ranking of Google Search Results

Neha V. Gawande[1], Reshma G. Sonone[2], Divya D. Jain[3], Pornima S. Sawade[4]

[1,2,3,4]*Student, Dept. of Computer Science & Engineering, DES'S COET, Dhamangaon Rly*

Abstract - *With more than two billion pages created by millions of Web page authors and organizations, the World Wide Web is a tremendously rich knowledge base. The knowledge comes not only from the content of the pages themselves, but also from the unique characteristics of the Web, such as its hyperlink structure and its diversity of content and languages. A considerably large portion of information present on the World Wide Web (WWW) today is in the form of unstructured or semi-structured text data bases. Copious material is available from the World Wide Web (WWW) in response to any user-provided query. It becomes tedious for the user to manually extract real required information from this material. Large document collections, such as those delivered by Internet search engines, are difficult and time-consuming for users to read and analyze. The detection of common and distinctive topics within a document set, together with the generation of multi-document summaries, can greatly ease the burden of information management. Information is essential to us in every possible way. We rely daily on information sources to accomplish a wide array of tasks. We need to sort out how to "organize" information. This work is an attempt to solve the problem of organizing information, specifically organizing web information. Because the largest information source today is the World Wide Web, and since we rely on this source daily for our tasks, it is of great interest to provide a solution for information categorization in the web domain. Clustering is useful technique in the field of textual data mining. Cluster analysis divides objects into meaningful groups based on similarity between objects. The work will focus on the problem of mining the useful information from the collected web documents using fuzzy clustering.*

***Key Words*: World Wide Web (WWW), Data mining, Clustering.**

1. INTRODUCTION

Different users with different needs submit queries with one or more keywords to web search engines through simple user interfaces. Search engines depend on keyword matching for searching against a collection of web pages to find the pages that would be returned. Therefore, current retrieval systems are not adaptive enough to satisfy user's search needs. Furthermore, some keywords could be ambiguous and have different meanings as in the search query "Ajax". For such query, users might have various goals and prefer different answers, i.e. "Ajax web based development", "Dutch football team Ajax Amsterdam", or "cleaning product Ajax". However, users often submit short queries in searching the web that does not provide adequate information to identify user needs.

The Web grows and evolves faster than we would like and expect, imposing scalability and relevance problems to Web search engines. There are three main data types in the Web: content (text, multimedia), structure (links that form a graph) and Web usage (transactions from Web logs). We emphasize the web mining. Server logs of search engines store traces of queries submitted by users, which include queries themselves along with Web pages selected in their answers. Query mining is based in the fact that user queries in search engines and Websites give valuable information on the interests of people. In addition, clicks after queries relate those interests to actual content

Many popular Web services rely on folksonomies including delicious (del.icio.us) and flickr (flickr.com). Despite the rising popularity of those Web services, research on folksonomies is still at an early stage. Much of the work has been focused on the study of the data properties, the analysis of usage patterns of tagging systems, the discovery of hidden semantics in tags, the using of annotations in enterprise search and the user's interest in discovery for personalized search. The objective of this paper, however, is to leverage the efforts and expertise of users embodied in social annotations for improving information retrieval.

Identification of pages of high quality and relevance to a query given by user is critical a goal of successful information retrieval on the web. There are different forms of web Information Retrieval that differentiate it and make it more challenging than previous problems occurred. The pages on the web contain links to other pages and it is possible to determine a more global notion of page quality by analyzing this web structure. The PageRank algorithm analyzes the entire web structure and provides the original basis for ranking in the Google search engine. Several other linked-based methods for ranking web pages have been proposed which includes both PageRank and HITS and in this area much more research is needed.

2. RELATED WORK

Web mining is the application of data mining techniques to extract knowledge from Web data, including Web documents, hyperlinks between documents, usage logs of web sites, etc. Web mining is the use of data mining techniques to automatically discover and extract information from Web documents and services. Even though it is strongly related to data mining, it is not equivalent to it. Three main axes of Web mining have been identified, according to the

Web data used as input in the data mining process, namely Web structure, Web content and Web usage mining.

The search engine then analyzes the contents of each page to determine how it should be indexed (for example, words can be extracted from the titles, page content, headings, or special fields called meta tags). Data about web pages are stored in an index database for use in later queries. A query from a user can be a single word. The index helps to find the information relating to the query as quickly as possible, some search engines, such as Google, store all or part of the source page (referred to as a cache) as well as information about the web pages, while others, like AltaVista, store every word of every page they find. This cached page always holds the actual search text since it is the one that is actually indexed, so it can be very useful when the content of the current page has been updated and the search terms are no longer in it.This problem might be considered a mild form of link rot or link death, and Google's handling of it increases usability by satisfying user expectations that the search terms will be on the returned webpage. This satisfies the principle of least astonishment, since the user normally expects that the search terms will be on the returned pages. Increased search relevance makes these cached pages very useful as they may contain data that may no longer be available elsewhere.

In this paper, an effective hybrid personalized re-ranking search approach is proposed by modeling user's search interests in a conceptual user profile, and then exploiting this profile in the re-ranking process. The user profile consists of concepts obtained by hierarchically classifying user's clicked search results into categories. These categories are extracted from a concept hierarchy called The Open Directory Project (ODP) where each concept represents a category. Any structural noise is removed from the ODP to obtain a more accurate concept hierarchy. Furthermore, each concept in the user profile consists of two types of documents; taxonomy document and viewed document. Taxonomy document is used to represent the user general interests as it contains information from web pages originally associated with such ODP category. Viewed document is used to represent the user specific interests as it contains information from web pages clicked by the user. Finally, the hybrid reranking process of search results is performed by semantically integrating user's general and specific interests from the user profile together with rankings of the traditional search engine.

3. IMPLEMENTATION

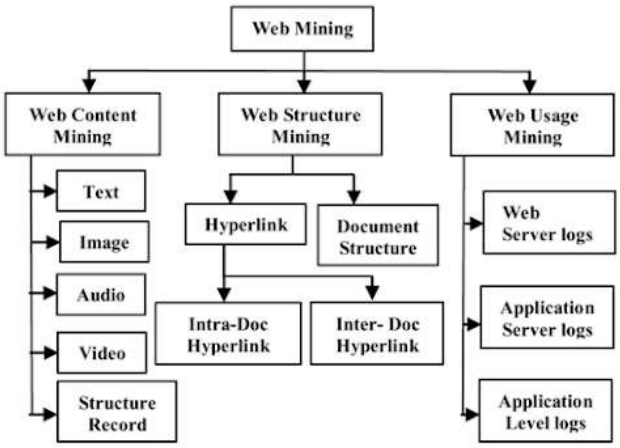

Fig.1 Conceptual Diagram

3.1 Web Mining

The patterns can be discovered from World Wide Web (www) using the technique called web mining. As the name suggests, the process of mining the web is done. It makes uses different approaches to identify and retrieve data from servers and organizations to get to both organized and unstructured information from browsers, server logs, websites and link structure, page content and different sources.

3.2 Web Content Mining

Web content mining has to do with the retrieval of information (content) available on the Web into more structured forms as well as its indexing for easy tracking information locations. Web content may be unstructured (plain text), semi-structured (HTML documents), or structured (extracted from databases into dynamic Web pages).

3.3 Web Usage Mining

The process of analyzing the user's browsing behavior is called Web usage mining. It can be regarded as a three phase process, consisting of the data preparation, pattern discovery and pattern analysis phases. In the first phase, Web data are preprocessed in order to identify users, sessions, page views, and so on. The input data are mainly the hits registered in the Web usage logs of the site, sometimes combined with other information such as registered user profiles, referrer's logs, cookies, etc.

3.4 Web Structure Mining

Web structure mining is based on the link structures with or without the description of links. Markov chain model can be used to categorize web pages and is useful to generate

information such as similarity and relationship between different websites. The structured summary about websites and web pages is the task which can be easily achieved through this web mining technique. It uses treelike structure to analyze and describe HTML or XML. Some algorithms have been proposed to model the Web topology such as HITS, PageRank and improvements of HITS by adding content information to the links structure and by using outlier filtering. These models are mainly applied as a method to calculate the quality rank or relevancy of each Web page. Some examples are the clever system and Google.

4. SCREENSHOTS

5. CONCLUSION

The results produced by the search engine are enormous and irrelevant to the user context. The proposed approach makes use of the text mining on the web pages listed by the search engine to rank. can be concluded that the implementation of the project would make it easy for the users to get their required data quickly and easily without requiring unnecessary searching. A critical goal of successful information retrieval on the web is fulfilled by identifying which pages are of high quality and relevance to a user's query. It will reduce the time taken for the suggested query and it's used to reduce the computational cost and improves

the classification accuracy. Finally it will retrieve the exact dataset for the suggested query.

REFERENCES

[1] O. Ben-Yitzhak, N. Golbandi, N. Har'El, R. Lempel, A. Neumann,S. Ofek-Koifman, D. Sheinwald, E. Shekita, B. Sznajder,and S. Yogev, "Beyond basic faceted search," in Proceedings of WSDM '08, 2008.

[2] M. Diao, S. Mukherjea, N. Rajput, and K. Srivastava, "Facetedsearch and browsing of audio content on spoken web," in Proceedings of CIKM '10, 2010.

[3] D. Dash, J. Rao, N. Megiddo, A. Ailamaki, and G. Lohman,"Dynamic faceted search for discovery-driven analysis," in CIKM '08, 2008.

[4] W. Kong and J. Allan, "Extending faceted search to the generalweb," in Proceedings of CIKM'14, ser. CIKM '14. New York,NY, USA: ACM, 2014, pp. 839–848.

[5] T. Cheng, X. Yan, and K. C.-C. Chang, "Supporting entity search: a large-scale prototype search engine," in Proceedings of SIGMOD '07, 2007, pp. 1144–1146.

[6] D. Sridevi et al., "Survey on Latest Trends in Web Mining", International Journal of Research in Advent Technologies, Vol. 2, No. 3, E-ISSN:2321-9637, March-2014.

[7] Gyanendra Kumar et al., "Page Ranking Based on Number of Visits of Links of Webpage", YMCA University of Science & Technology, Faridabad, India(ICCCT-2011).

[8] Sungrim Kim, "Information Retrieval using Context Information on the Web 2.0 Environment", Joonhee Kwon (IJCSNS2009).

[9] Bing Liu and Kevin Chen Chuan Chang, "Editorial: Special Issue onWeb Content Mining", SIGKDD Explorations, Volume 6, Issue 2.

[10] Bin W andLiuZhijing, "Web Mining Research", 5th InternationalConference on computational Intelligence and MultimediaApplications, 2003.

[11] G. Poonkuzhali et al., "Signed Approach for Mining Web Content Outliers", International Science Index waset.org/Publication/13636Vol:3, No:8, 2009.