Re-Ranking of Google Search Results using Data Mining

Ms. Manali Kohale[1] Ms. Aishwarya Ghogre[2] Ms. Shweta Barbade[3] Ms. Snehal Nagre[4]
Prof. Ms. P. B. Lohiya[5]
[1,2,3,4,5]PRMITR, Badnera, India

Abstract— For finding information on the WWW, Search has possibly become the dominant paradigm. We encounter a number of problems of finding information on the web. In this report, we discuss approaches that have been employed in various research programs such as Google. We show that leveraging the vast amounts of data on web, it is possible to successfully address problems in innovative ways that vastly improve on standard, but often data impoverished, methods. Similar material is available from the World Wide Web in response to any keyword searched by user. Extracting real required information manually from this material becomes difficult for the user. The load of information management can be simplified by the detection of topics which can be common and distinctive within a document set, together with the generation of multi-document summaries. This work is an attempt to get efficient and accurate result of the query provided by user in web search engine.
Key words: Data Mining, Search Engine

I. INTRODUCTION

The information on the World Wide Web is searched using web search engine. The search engine results pages are the lines of result generated by web search engine. The information is a mix of web pages, images, and other types of files. The way of presenting, storing, organizing and accessing the information items is called Information Retrieval. The representation and organization of information should be in such a way that the user can access information to meet his information need. This project is attempt to create an application using web mining techniques like content mining, usage mining and structure mining to give an efficient result of a search.

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 [1], 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 [3][4], and in this area much more research is needed.

A. Objective

This work will try to achieve some or all of the following objectives.

- To collect the web pages related to application domain.
- To generate various rules as per selected domain.
- To implement fuzzy clustering in web text mining.
- To retrieve (mine) relevant information to the user from collected web pages.
- To analyze the retrieved result

B. Motivation

The existing Web information retrieval contains various problems which can be solved by data mining techniques. In this report, we have presented a number of challenges, giving an overview of some approaches taken for solving these problems and for promoting future work. As a result, we hope to encourage more research in this area. Thus by implementing various data mining technique will help to achieve achieve the goal of organizing the web information and making it efficient and easily accessible.

II. LITERATURE REVIEW

Eugene Agichtein et al. [1] proposed Improving Web Search Ranking by Incorporating User Behavior Information and it incorporating implicit feedback can augment other features, improving the accuracy of a competitive web search ranking algorithms. Author explored the utility of incorporating noisy implicit feedback obtained in a real web search setting to improve web search ranking.

M. J. Cafarella, [5] The World-Wide Web consists of a huge number of unstructured documents, but it also contains structured data in the form of HTML tables. We extracted 14.1 billion HTML tables from Google's general-purpose web crawl, and used statistical classification techniques to find the estimated 154M that contain high-quality relational data. Each relational table having its own schema of labeled and typed columns, can be considered a structured database. The eff ective techniques are used for searching for structured data at search-engine scales.

Weize Kong, [9] Peresent Faceted search facilitates users to search for a multi-dimensional information space by getting together query searched with drill-down options in each and every facets. For example, when searching "computer monitor" in an e-commerce site, users can select brands and monitor types from the provided facets {"Samsung", "Dell", "Acer",} and {"LET-Lit", "LCD", "OLED" ...}. It has been used for many applications like e-commerce and digital libraries. We present both intrinsic evaluation, which evaluates facet generation on its own, and extrinsic evaluation, which evaluates an entire Faceted Web Search system by its utility in assisting search clarification. We also design a method for building reusable test collections for such evaluations. Our experiments show that using the Faceted Web Search interface can significantly improve the original ranking if allowed sufficient time on user feedback on facets.

The list of concerned queries is proposed by R. Baeza-Yates [11], which are founded in antecedently published queries, and thus the required queries can be published by the user to redirect the search process. The method proposed is based on a query clustering procedure in which groups of semantically like queries are named. The clustering procedure uses the content of historical preferences of users registered in the query log of the search engine. This

method ranks the queries according to a relevance criterion and also discloses the related queries. As future work we tend to improve the notion of interest of the recommended queries and to develop alternative notions of interest for the question recommender system.

III. PROPOSED SYSTEM

The proposed system will apply three different mining techniques such as web structure mining,web content mining and web usage mining on the Googles links and will rearrange the links in more efficient manner.

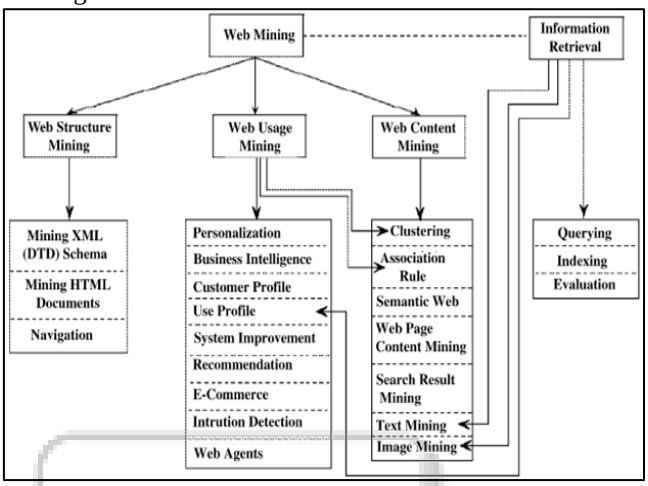

Fig. 1: Conceptual Diagram

A. 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.

The three general sets of information: previous usage patterns, degree of shared content [5] and inter-memory associative link structures [26] corresponding to the three subsets in Web mining namely:

1) Web usage mining,
2) Web content mining and
3) Web structure mining.

These three forms the base of Web mining analysis.

B. Web Usage Mining

The Web usage mining is also known as Web Log mining, which is used to analyze the behavior of website users. This focuses on technique that can be used to predict the user behavior while user interacts with the web. It also uses the secondary data on the web where the activity involves automatic discovery of user access patterns from one or more web servers. It contains four processing stages including data collection, preprocessing, pattern discovery and analysis [102].Some algorithms been proposed for web usage mining are FP Growth and Apriori algorithm.

C. 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 [14], PageRank [23] and improvements of HITS by adding content information to the links structure [7] and by using outlier filtering [22]. 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 [7] and Google [16].

D. Web Content Mining

The identification of useful information from web contents which include text, image, audio, video, etc is done using technique called web content mining. The mining of link structure aims at developing techniques to take advantage of the collective judgment of web page quality which is available in the form of hyperlinks that is web structure mining [27]. It includes extraction of structured data/information from web pages, identification, similarity and integration of data's with similar meaning, view extraction from online sources, and concept hierarchy, knowledge incorporation [1].One of the popular algorithms for web content mining is k means algorithm.

IV. CONCLUSION

There has been constant efforts in web mining techniques to come out with most efficient web searching methods for retrieving useful information from the web pages. Web Structure Mining, Web Usage Mining and Web Content Mining play a vital role in achieving this. In this project we propose a new architecture which is a blend of these three techniques. Three algorithms Aprori Algorithm, K-Means Algorithm and HITS Algorithm will be implemented and results will be evaluated for different cases. From the obtained results it will be evident that these algorithms explore most relevant pages on the top of search results.

It 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.

In similar line other existing algorithms could be analyzed for efficient Information retrieval.

Thus the use of concept based mining algorithms for content mining, structure mining and usage mining and content based filtering techniques to retrieve the exact data for the suggested query of the web user from the web server will help the web user to satisfy their needs and concise the web search time. 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] K. Balog, E. Meij, and M. de Rijke, "Entity search: building bridges between two worlds," in Proceedings of SEMSEARCH'10, 2010, pp. 9:1–9:5.1041-4347 (c) 2015 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.Seehttp://www.ieee.org/publications_standards/publications/rights/index.html for more information. This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TKDE.2015.2475735, IEEE Transactions on Knowledge and Data Engineering IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING 14
[7] M. Bron, K. Balog, and M. de Rijke, "Ranking related entities:components and analyses," in Proceedings of CIKM '10, 2010,pp. 1079–1088.
[8] C. Li, N. Yan, S. B. Roy, L. Lisham, and G. Das, "Facetedpedia: dynamic generation of query-dependent faceted interfaces forwikipedia," in Proceedings of WWW '10. ACM, 2010.
[9] W. Dakka and P. G. Ipeirotis, "Automatic extraction of useful facet hierarchies from text databases," in Proceedings of ICDE'08, 2008, pp. 466–475.
[10] Herdagdelen, M. Ciaramita, D. Mahler, M. Holmqvist,K. Hall, S. Riezler, and E. Alfonseca, "Generalized syntacticand semantic models of query reformulation," in Proceeding of SIGIR'10, 2010.

