IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

# Re-Ranking Of Google Results Based On Information Mining Techniques*

*Gayatri Vilas Bagade*[1]
*Department of Computer Science*
G H Raisoni University,
Amravati, India

*Prof. Dr Rais Abdul Hamid Khan*[2]
Assist prof, Department of Computer Science and engineering
G H Raisoni University,
Amravati, India

***Abstract*****—***In Re-ranking of Google result will provide you the enhance way for searching the information on any web site by using content mining techniques. The World Wide Web is a system of interlinked hypertext documents that are accessed via the internet. It plays a leading role for retrieving user requested information from the web resources. In order to retrieve user requested information, search engine plays a major role for crawling web content on different node and organizing them into result pages so that user can easily select the required information by navigating through the result pages link. This strategy worked well in earlier because, number of resources available for user request is limited. It is feasible to identify the relevant information directly by the user from the search engine results. As the Internet era increases, sharing of resource also increases and this leads to develop an automated technique to rank each web content resource. Different search engine uses different techniques to rank search results for the user query. This leads to business motivation of bringing up their web resource into top ranking position. As the competition and web resource increases, the ranking of web content becomes tedious and dynamic with respect to the user query.*

***Keywords—*** *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 Page Rank 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 Page Rank and HITS 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.

**C.** Data Mining

Web usage mining is a subset of web mining operations which itself is a subset of data mining in general. The aim is to use the data and information extracted in web systems in order to reach knowledge of the system itself. Data mining is different from information extraction although they are closely related. To better understand the concepts brief definitions of keywords can be given as [1]. Data:- "A class of information objects, made up of• units of binary code that are intended to be stored, processed, and transmitted by digital computers". Information:-"is a set of facts with processing• capability added, such as context, relationships to other facts about the same or related objects, implying an increased usefulness. Information provides meaning to data" Knowledge :- "is the summation of information• into independent concepts and rules that can explain relationships or predict outcomes" Information extraction is the process of extraction information from data sources whether they are structured, unstructured or semi-structured into structured and computer understandable data formats. Area where data mining is widely used is bioinformatics where very large data about protein structures, networks and genetic material is analyzed. The sub category of interest in this thesis is the web mining which acts on the data made available in the World Wide Web (WWW) data servers.

D. Web Content Mining

"Web content mining describes the automatic search of information resources available on-line". The focus is on the content of web pages themselves. content mining as agent based approaches; where intelligent web agents such as crawlers autonomously crawl the web and classify data and database approaches; where information retrieval tasks are employed to store web data

in databases where data mining process can take place Most web content mining studies have focused on textual and graphical data since the early years of internet mostly featured textual or graphical information. Recent studies started to focus on visual and aural data such as sound and video content too

## II. LITERATURE SURVEY

In Year 2011, D. Choi has defined an approach to perform the query over the web and to extract the web document. The author also presented the approach to assign the ranking to these web documents. With the development of web search engines, one of the major tasks is to retrieve these documents from web effectively. These search engines uses the some ranking algorithm to present the result in an effective way.

The author has defined a study of existing ranking algorithm used by different search engines. The author explored the advantages and limitations of these ranking algorithms. The major contribution of author was the definition of query based information retrieval. The author defined the classification over a query and performed the query filtration. Based on this analysis, the ranking is improved and refined

Zhou Hui has presented a work on optimization of search engine under the keyword analysis along with face link analysis and back link analysis. Author defined a relational environment based on search engine optimization of that the search ranking will be improved. Author also discussed various aspects of search engine optimization including the optimization vector, ranking, working principal etc

. Ping-Tsai Chun has presented a search engine optimization approach under the current market scenario analysis. Author defined the web service analysis to improve the business dictating and to provide the work under small organizations so that the effective keyword analysis based search will be performed. Author presented the work for text search as well as for image search over the web. The pattern analysis is defined to perform the effective search over web.

One of the common model for web page ranking and prediction system is defined by Markov Model. Such model defines the navigational behavior of web graph theory as well as defines the transitional probabilities over the ranking analysis. The author not only defined work for a single web page access, but also presents the work for web path generation. The web path is actually defined as a series of web pages that a user can visit after visiting a specific web page. To perform this kind of analysis a Markov Model based prediction system is defined. The prediction is here defined under the web usage mining that defined the structural information for prediction of web pages. To perform such kind of analysis, the author defined a web page graph and implements the markov model over it to analyze the frequency match. Based on which a acyclic web path is generated and based on the weightage assigned to this web path the prediction is performed

Another work in web page ranking is the comparison of different web pages and the web sites. Author M. Klein performed this comparison on two football team web sites of college team. The analysis is performed under the web page metrics to perform the quality assessment. The author has defined the page comparison and the ranking system under the graph theory

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 generalpurpose web crawl, and used statistical classification techniques to find the estimated 154M that contain highquality 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

## III. 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.

## IV. 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_standa rds/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.