International Journal Of

Recent Scientific Research

ISSN: 0976-3031
Volume: 7(5) May -2016

ENHANCING PERFORMANCE OF INFORMATION RETRIEVAL IN WEB SEARCH ENGINE RESULTS USING INTEGRATED APPROACH OF WEB MINING

Keole R.R., Karde P.P and Thakare V.M

THE OFFICIAL PUBLICATION OF
INTERNATIONAL JOURNAL OF RECENT SCIENTIFIC RESEARCH (IJRSR)
http://www.recentscientific.com/ recentscientific@gmail.com

ISSN: 0976-3031

Available Online at http://www.recentscientific.com

International Journal of Recent Scientific Research
Vol. 7, Issue, 5, pp. 11039-11043, May, 2016

International Journal of Recent Scientific Research

Research Article

ENHANCING PERFORMANCE OF INFORMATION RETRIEVAL IN WEB SEARCH ENGINE RESULTS USING INTEGRATED APPROACH OF WEB MINING

Keole R.R[1]., Karde P.P[2] and Thakare V.M[3]

[1]Department of Information Technology, H. V. P. M. College of Engineering & Technology, Amravati, India
[2]Department of Information Technology, Govt. Women's Residential Polytechnic, Yavatmal, India
[3]Department of Computer Science, Sant Gadge Baba Amravati University, Amravati, India

ARTICLE INFO

Article History:

Received 29th February, 2016
Received in revised form 19th March, 2016
Accepted 25th April, 2016
Published online 28th May, 2016

Keywords:

Web Content Mining, Web Structure mining, Web Usage Mining, and Web Search Engine.

ABSTRACT

The World Wide Web (WWW) is popular and interactive medium to propagate information today. When a user makes a query from search engine, it generally returns a large number of pages in response to user queries. This result-list contains many relevant and irrelevant pages according to user's query. As user impose more number of irrelevant pages in the search result-list, to assist the users to navigate in the result list, various ranking methods are applied on the search results. The search engine uses these ranking methods to sort the results to be displayed to the user. Information Retrieval is the activity of obtaining information resources relevant to an information need from a collection of information resources. Therefore in this paper an Approach to integrate web content, Web structure mining which takes into account the hyperlinks of the web pages & Web server log files to discover useful information is proposed. The work will focus on the problem of improving the performance of information retrieval in web search engine results.

INTRODUCTION

The World Wide Web contains several billions of information and is still growing at a very faster rate as most of the people use the internet for retrieving interesting document. But most of the time, they lose their temper by getting lot of insignificant document even after navigating several links. Thus developing user friendly tool for retrieving the relevant content without accessing the complete data on the outset has become an important concern among the Web mining research communities [1, 2]. Searching is considered a very important aspect of the Internet. In the age of Google, Yahoo!, Bing and others where each one of them is trying to outdo the other in terms of performance for their search engines, it is apparent that search and its related technologies is an important research area. There are tens and hundreds of search engines available but some are popular like Google, Yahoo, Bing etc., because of their crawling and ranking methodologies. The search engines download, index and store hundreds of millions of web pages. They answer tens of millions of queries every day. So Web mining and ranking mechanism becomes very important for effective information retrieval [26]. The searching usually involves searching over large content on the web [25]. In today's age, where the amount of information available on the Internet is millions of Giga-bytes, it becomes highly necessary to implement efficient search techniques in order to index and rank such massive amounts of data. There are different aspects/parts which are involved in implementing a successful and efficient search engine [9, 10, 16]. The different steps involved are: Crawling the Internet for all the data, indexing all the data according to a certain efficient model, Ranking of these indexed documents to give a clear demarcation between the documents which are more frequently viewed to the ones that are not and displaying the appropriate results. There are three kinds of information that have to be dealt with when any user is accessing any web site [3,11,13]. So the three types of information are based on content of data, structure of data and log data. Based on these three types information, research area of web mining has been divided into web content mining, web structure mining and web usage mining [2,13]. Web content

**Corresponding author: Keole R.R*
Department of Information Technology, H. V. P. M. College of Engineering & Technology, Amravati, India

mining is the process of extracting knowledge from the content of documents or their descriptions. Web structure mining is the process of inferring knowledge from the World Wide Web organization and links between references and referents in the Web. Technically Web usage mining is the process of extracting useful information from web server logs i.e. user's history [3, 15] also known as Web Log Mining [4].

Related Work

World Wide Web has become a powerful platform to store and retrieve information as well as mine useful knowledge. Due to the properties of the huge, diverse, dynamic and unstructured nature of Web data, Web data research has encountered a lot of challenges. Characteristics of web and various issues on web content mining are presented in [1].

R. Kosala & H. Blockeel presented a survey related to the research in the area of Web mining; they focused on the term Web mining and suggested three Web mining categories.

In [6] Z. Lu & H. Zha States those different users may have different search goals when they submit a query to a search engine. The inference and analysis of user search goals can be very useful in improving search engine relevance and user experience. Here a novel approach to infer user search goals by analyzing search engine query logs was proposed.

Ida Mele [7] proposed that Data, stored in server logs, represents a valuable source of information. The research focuses on two important issues: improving search-engine performance through static caching of search results, and helping users to find interesting web pages by recommending news articles and blog posts.

A novel approach using weighted technique [9] is introduced to mine the web contents catering to the user needs. Experimental results prove that the performance of the proposed approach in terms of precision, recall and F-measure is high when compared to other search engine results. Algorithm used is Relevancy and Weight based approach.

A new method is presented in [14] to identify navigation related Web usability problems based on comparing actual and anticipated usage patterns. The actual usage patterns can be extracted from Web server logs and then applying a usage mining algorithm to discover patterns among actual usage paths.

W. Xing and A.Ghorbani [17] proposed a Weighted Page Rank (WPR) algorithm which is an extension of the Page Rank algorithm. This algorithm assigns a larger rank values to the more important pages rather than dividing the rank value of a page evenly among its outgoing linked pages. Each outgoing link gets a value proportional to its importance. The importance is assigned in terms of weight values to the incoming and outgoing links.

In [18] S. Nimgaonkar & S. Duppala presented detailed study about web Mining and web content mining also present a comprehensive survey of some of the techniques of web content mining used in real time for the extraction of structured and semi-structured data.

Hao Chen and Susan Dumais [22] developed a user interface that organizes web search results into hierarchical categories. Automatic text classification technique was used to classify arbitrary search results.

T. Joachims [23] proposed an approach to automatically optimizing the retrieval quality of search engine using click-through data stored in query logs and the log of links the users clicked on in presented ranking.

R.Bhushan and R.Nath [24] presented a web recommendation approach based on learning from web logs and recommends user a list of pages which are relevant to him by comparing with user's historic pattern. Search result list is optimized by re-ranking the result pages.

Web Mining Approach

Oren Etzioni [11] was the person who coined the term Web Mining first time. Web mining is also a cross point of database, information retrieval and artificial intelligence. The most common way of representing text documents is using the Vector Space Model (VSM) [28], where each document is represented as a feature vector, which length corresponds to the number of unique attributes used for representing documents in the collection. Each vector component, that is, each feature, has an associated weight which indicates the importance of that attribute to characterize or represent the document. Web mining is the Data Mining technique that automatically discovers or extracts the information from web documents.

Web Mining Taxonomy

Web Mining can be broadly divided into three distinct categories, according to the kinds of data to be mined:

Web Content Mining

Web content mining is the process of extracting useful information from the contents of web documents. Web Content Mining (WCM) is concerned with the retrieval of information from WWW into more structured form and indexing the information to retrieve it quickly [9, 18]. These web documents are collection of images, audio, video, text and structured records (such as tables and lists). Web content mining has the following approaches to mine data (1) Unstructured text mining, (2) structured mining, (3) Semi structured text mining, and (4) Multimedia mining [35].

Web Structure Mining

Web structure mining generates structured summaries about information on web pages. It shows the links from one web page to the other web page, known as hyper link [17, 37]. A hyperlink is a structural unit that connects a location in a web page to a different location, either within the same web page or on a different web page. Intra-document web pages connect different parts of web pages using hyper link. Inter-document web pages connect two different web pages using hyperlink [37]. The popularity of the web page is generally measured by the fact that a particular page should be referred by large number of other pages and the importance of web pages may be adjudged by a large number of out links contained by a page.

Web Usage Mining

Web Usage Mining (WUM) is responsible for recording the user profile and user behavior inside the log file of the web. Web usage mining process is used to extract useful information from the data which is derived by the user while surfing on the Web [4, 6, 7]. It extracts data stored in server access logs, referrer logs, agent logs, client-side cookies, user profile and meta-data.

EXISTING METHODOLOGY OF INFORMATION RETRIEVAL USING WEB SEARCH ENGINE

An information retrieval process begins when a user enters a query into the system [25, 36]. To retrieve required information from World Wide Web, search engines perform number of tasks based on their respective architecture [4, 5]. The web search engine represents the user interface needed to permit the user to query the information. It is the connection between user and the information repository when user sends query to search engine and returns a list of documents where the keywords were found. The search results are generally presented in a line of results often referred to as Search Engine Result Pages (SERPs). There are 3 important components in a search engine. They are Crawler, Indexer and Ranking mechanism. The crawler is also called as a robot or spider that traverses the web and downloads the web pages which are sent to an indexing module that parses the web pages and builds the index based on the keywords in those pages. When a user types a query using keywords on the interface of a search engine, the query processor component match the query keywords with the index and returns the URLs of the pages to the user. But before showing the pages to the user, a ranking mechanism is done by the search engines to show the most relevant pages at the top and less relevant ones at the bottom. The search engines consider two main areas when determining what the website is about and how to prioritize it [36].

1. Content on the website: Information about what topics the website specializes in and scanning its code for certain tags, descriptions and instructions.
2. who's linking to the site: The more inbound links a website has, the more influence or authority it has. Essentially, every inbound link counts as an increase in priority for that websites content. Also, each inbound link holds different weight [30].

The Search Engine Results Page (SERP) is then listed in order of most relevant. If the same search is conducted on different search engines, chances are that different results could be seen on the SERP. This is because each search engine uses their own algorithm that considers multiple factors in order to determine what results to show in the SERP when a search query is entered. A few factors that a search engine algorithm may consider when deciding what information to show in the SERP include:

- Geographic location of the searcher & its Historical performance (clicks, etc.)
- Link quality & Link type
- Webpage content (keywords, tags, pictures) & Back end code or HTML of webpage.

Issues Related To Information Retrieval From Web

Keeping information organized is an important issue to make information retrieval easier. Although the information we need is sometimes available on the Web, this information is only useful if we have the ability to find it. Current search engines return lists of ranked url's with their title and their snippet (a short description of the document), but still fail to find relevant contents and to present them in an organized way, therefore the user is required to go through the extensive list of the retrieved results to satisfy its needs[11]. Most search engines perform in a query-triggered way that is mainly on a basis of one keyword or several keywords entered. Sometimes the results returned by the search engine don't exactly match what a user really needs due to the fact of the existence of the homology. To improve the Internet service quality and increase the user click rate on a specific website, thus, it is necessary for a Web developer or designer to know what the user really wants to do, predict which pages the user is potentially interested in, and present the customized Web pages to the user by learning user navigational pattern knowledge. Therefore, the main problem is regarding the retrieval of relevant web pages. One of the Solutions to this is the combine approach of Web Usage Mining, Web Content Mining and Web Structure Mining for Enhancing Search-Result Delivery. Therefore the main aim of the proposed research work is to design an efficient approach to improve the performance of information retrieval in web search engine results [40] that is able to reorder the web documents effectively [11].

Architecture Design of the Proposed Methodology

Most of the research work is focused only on web usage mining, web content mining or web Structure mining for Enhancing Search-Result Delivery, for improving the performance of Information Retrieval in web search engine results. Combine approach of Web Usage Mining, Web Content Mining and Web Structure Mining for Enhancing Search-Result delivery is not considered. In this research work emphasis will be given on information retrieval based on the combine approach of Web content– free text, Web structure–hyperlinks, and Web usage-web log data.This is to increase the accuracy factor and relevancy in retrieval of information from web. For Web Content mining a term-based weighted technique will be used to mine the web contents which will improve the performance of web search engine results in terms of precision, recall and F-measure notations. In web structure mining, Weighted Page Rank algorithm (WPR) takes into account for the importance of both the in-links and the out-links of the pages which distributes rank scores based on the popularity of the pages. Web Logs are important for information repositories, which will record the user activities on search results to infer user search goals by analyzing search engine query logs. This technique provides a framework to discover different user search goals for a query by clustering the proposed feedback sessions. Feedback sessions are constructed from user click-through logs and can efficiently reflect the information needs of users. The Search result is optimized by re-ranking the search result pages. This proposed system proves to be efficient as the pages desired by the user will be on the top priority in the search result list.

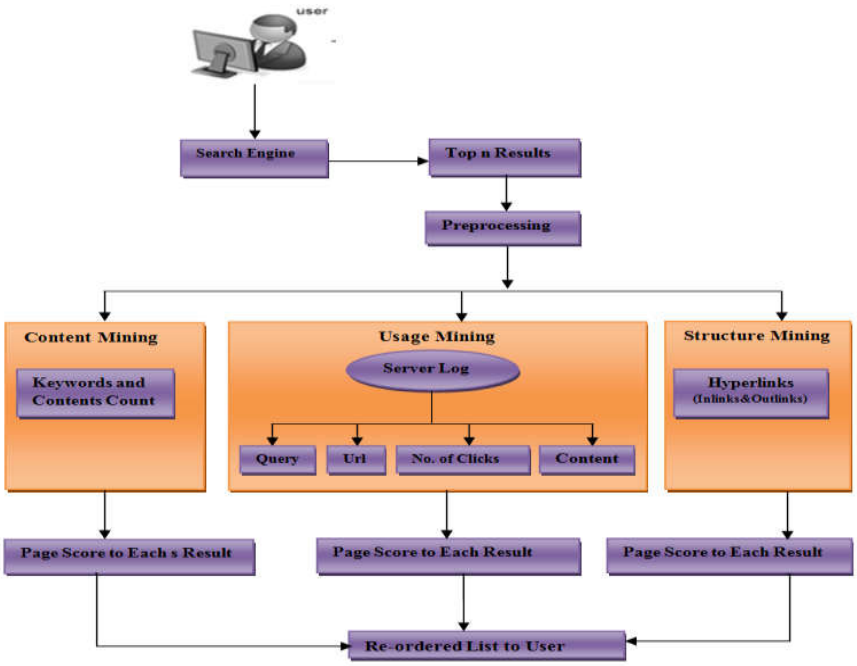

Figure Architecture of Proposed System

Hence the research will carry out the work to design an efficient approach to improve the performance of information retrieval in web search engine results. The proposed methodology can be incorporated within a web based search engine to provide better performance.

CONCLUSION

Web mining is a growing research area in the mining community. Retrieving relevant content from the web is a very common task. However, the results obtained, by most of the search engines do not necessarily produce result that is best possible catering to the user needs. To extract the specific data from web, the three categories Web Content Mining, Web Structure Mining and Web Usage Mining of web mining play a major role. The proposed methodology, in this paper focuses on the integrated approach of Web content– free text, Web structure–hyperlinks, and Web usage-web log data to improve the performance of information retrieval in web search engine results; this work will focus on mining of the useful information as per the user query from the web documents. Finally the Search result is optimized by re-ranking the search result pages. This proposed system proves to be efficient as the pages desired by the user will be on the top priority in the search result list.

References

1. Bing Liu, Kevin Chen- Chuan Chang, and Editorial: Special issue on Web Content Mining, *SIGKDD Explorations*, Volume 6, and Issue 2.
2. G.Poonkuzhali, K.Thiagarajan, K.Sarukesi and G.V. Uma, SignedApproach for Mining Web content Outliers, *Proceedings of World Academy of Science , Engineering and Technology*, Vol.56,2009,PP 820-824.
3. Raymond Kosala, Hendrik Blockeel, Web Mining Research: A Survey, *ACM SIGKDD*, July 2000.
4. R. Baeza-Yates, C. Hurtado, and M. Mendoza, "Query Recommendation Using Query Logs in Search Engines," Proc. Int'l Conf. Current Trends in Database Technology (EDBT '04), pp. 588-596, 2004.
5. H. Chen and S. Dumais, "Bringing Order to the Web: Automatically Categorizing Search Results," Proc. SIGCHI Conf. Human Factors in Computing Systems (SIGCHI '00), pp. 145-152, 2000.
6. Z. Lu, H. Zha, X. Yang, W. Lin, Z. Zheng, "A New Algorithm for Inferring User Search Goals with Feedback Sessions," Proc. IEEE Transactions on Knowledge and Data Engineering, pp. 502-513, 2013.
7. Mele, " Web Usage Mining for Enhancing Search – Result Delivery and Helping Users to Find Interesting Web Content," ACM SIGIR Conf. Research and Development in Information Retrival (SIGIR '13), pp. 765-769, 2013.
8. J. Srivastava, R. Cooley, M. Deshpande, and P.-N. Tan. Web usage mining: discovery and applications of usage patterns from Web data. SIGKDD Explor. Newsl., 1(2):12, 2000.
9. P. Sudhakar, G. Poonkuzhali, R. Kishor Kumar, "Content Based Ranking for Search Engines," Proc. International Multi Conference of Engineers and Computer Scientists (IMECS 12), 2012.
10. X. Wang & C.-X Zhai, "Learn from Web Search Logs to Organize Search Results", Proc. 30th Ann. Int. ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR '07), pp. 87-94, 2007.
11. O. Zamir and O. Etzioni, "Web Document Clustering: A Feasibility demonstration," ACM (SIGIR, 99), pp. 46-54., 1998.
12. Guandong Xu," Web Mining Techniques for Recommendation and Personalization", Ph.D. dissertation, Victoria University, Australia, March 2008.

13. Ramakrishna, M.T. Gowdar, L.K. Havanur, M.S. Swamy (2010), "Web Mining: Key Accomplishments, Applications and Future Directions", *International Conference on Data Storage and Data Engineering (DSDE), pp.187 – 191,* 2010.
14. Ruili Geng*, Member, IEEE*, and Jeff Tian*, Member, IEEE*," Improving Web Navigation Usability by Comparing Actual and Anticipated Usage ",IEEE Transactions On Human-Machine Systems, Vol. 45, No. 1, February 2015.
15. Wang Bin and Liu Zhijing, "Web Mining Research", *in Proceeding of the 5th International Conference on Computational Intelligence and Multimedia Applications (ICCIMA'03)* 2003.
16. Brin, S. and L. Page, *The anatomy of a large-scale hypertextual Web search engine.* Comput. Netw. ISDN Syst., 1998. **30**(1-7): p. 107-117.
17. Wenpu Xing and Ali Ghorbani," Weighted Page Rank Algorithm", IEEE, 2004.
18. Satyajeet Nimgaonkar and Suryaprakash Duppala," A Survey on Web Content Mining and extraction of Structured and Semi structured data", ACM, 2012.
19. J. Srivastava, R. Cooley, M. Deshpande, and P.-N. Tan. Web usage mining: discovery and applications of usage patterns from Web data. SIGKDD Explor. Newsl., 1(2):12{23, 2000.
20. X. Wang and C.-X Zhai, Learn from Web Search Logs to Organize Search Results, Proc. 30th Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR '07), pp. 87-94, 2007.
21. Hua-Jun Zeng, Qi-Cai He, Zheng Chen, Wei-Ying Ma, Jinwen Ma," Learning to Cluster Web Search Results", ACM, 2004.
22. Hao Chen and Susan Dumais," Bringing Order to the Web: Automatically Categorizing Search Results", ACM, 2012.
23. T. Joachims, *"Optimizing Search Engines Using Clickthrough Data,"* Proc. Eighth ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (SIGKDD '02), pp. 133-142, 2002.
24. Ravi Bhushan and Rajender Nath, "Recommendation of Optimized Web Pages to Users Using Web Log Mining Techniques", Published by the IEEE Computer Society, IEEE 2012.
25. C. D. Manning, P. Raghavan, and H. Schtze "Introduction to Information Re-trieval", Cambridge University Press, 2011.
26. Duhan, N., A.K. Sharma, and K.K. Bhatia. Page Ranking Algorithms: A Survey. In Advance Computing Conference, 2009. IACC 2009. IEEE International. 2009.
27. Hsinchun Chen and Michael Chau, "Web Mining: Machine learning for Web Applications", Annual Review of Information Science and Technology 2003.
28. Salton, G., Wong, A., Yang, C.S. (1975). A vector space model for automatic indexing. *Communications of the ACM*, 18(11):613-620, 1975.
29. Nicholas O. Andrews and Edward A. Fox, "Recent Development in Document Clustering Techniques", Dept of Computer Science, Virgina Tech 2007.
30. Xu, J. and Li, H. " AdaRank: A Boosting Algorithm for Information Retrieval, Proceedings of the 30th Annual International ACM SIGIR Conference, Amster-dam, Netherlands, 2011.
31. RonGiles, How Search Engines Work, Available: http://www.website-consultant.co.nz/ Website/Top+10+Search+Engine+Ranking+Factors/How+Search+Engines+work.html.
32. H. Cunningham, N. Fuhr, and B. Stein "Challenges in Document Mining, Dagstuhl Seminar Proceedings, vol.1, no.4, pp.65-99, Germany, 2011.
33. Thomas Mandl Recent Developments in the Evaluation of Information Retrieval Systems: Moving Towards Diversity and Practical Relevance, Proceedings of the 22nd ICML, Finland, 2007.
34. V. Fresno and A. Ribeiro. An Analytical Approach to Concept Extraction in HTML Environments. Journal of Intelligent Information Systems - JIIS. Kluwer Academic Publishers, 215-235, 2004.
35. Johnson, F., Gupta, S.K., *Web Content Minings Techniques: A Survey*, *International Journal of Computer Application. Volume* 47 – No.11, p44, June (2012).
36. Laxmi Choudhary and B. Shankar Burdak,"Role of Ranking Algorithms for Information Retrieval", *International Journal of Artificial Intelligence and Applications* (IJAIA), Vol.3, No.4, July 2012.
37. T.Munibalaji, C.Balamurugan, ―Analysis of Link Algorithms for Web Mining, *International Journal of Engineering and Innovative Technology* (IJEIT), ISSN: 2277-3754, Volume 1, Issue 2, February 2012, pp-81-86.
38. Garza Villarreal, S. E., Martínez Elizalde, L., and Canseco Viveros, A. Clustering hyperlinks for topic extraction: An exploratory analysis. In Proceedings of the 2009 Eighth Mexican International Conference on Artificial Intelligence, MICAI '09, pages 128–133, Washington, DC, USA, 2009. IEEE Computer Society. ISBN 978-0-7695-3933-1, 2009.

****

How to cite this article:

Keole R.R., Karde P.P and Thakare V.M.2016, Enhancing Performance of Information Retrieval in web Search Engine Results using Integrated Approach of Web Mining. *Int J Recent Sci Res.* 7(5), pp. 11039-11043.

TSSN 0976-3031