INTERNATIONAL JOURNAL OF PURE AND APPLIED RESEARCH IN ENGINEERING AND TECHNOLOGY

A PATH FOR HORIZING YOUR INNOVATIVE WORK

AN EFFICIENT APPROACH TO HYBRIDIZE WEB CONTENT, WEB STRUCTURE AND WEB USAGE MINING FOR ENHANCING WEB SEARCH ENGINE RESULTS

MR. RANJIT R. KEOLE[1], DR. P. P. KARDE[2]

1. Research Scholar, Dept. of Information Technology, H.V.P.M's COET, Amravati, India.
2. HOD, Dept. of Information Technology, H.V.P.M's COET, Amravati, India.

Accepted Date: 15/03/2016; Published Date: 01/05/2016

Abstract*:* Search engine has become an important tool in today's world for searching various data but while searching many users end up with irrelevant information causing a waste in user time and accessing time of the search engine. So to narrow down this problem, many researchers are involved in web mining. Web mining is universal set of Web Structure Mining, Web Usage Mining and Web content Mining. In present scenario web mining is the most active area where the research is going on rapidly. According to literature review most of the research work is focused either on web content, web structure or web usage mining for Enhancing Search Result Delivery. Combine approach of Web Usage, Web Content and Web Structure Mining is not considered for improving the performance of Information Retrieval in web search engine results. In this paper we are proposing an Approach to hybridize web content, web structure & web usage mining for Enhancing Web Search Engine Results Delivery. Finally, the Search result is optimized by re-ranking the result pages.

Keywords: Web Content Mining, Web Structure Mining, Web Usage Mining, Search Engine Results.

Corresponding Author: MR. RANJIT R. KEOLE

Access Online On:

www.ijpret.com

How to Cite This Article:

Ranjit R. Keole, IJPRET, 2016; Volume 4 (9): 604-615

PAPER-QR CODE

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].

Web users are always drowning in an "ocean" of information and facing the problem of information overload when interacting with the web. Typically, the following problems are often mentioned in Web related research and applications [12]:

1. Finding relevant information: To find specific information on the web, users use a search engine as a search assistant, for which he or she often enters one or several keywords as a query, then the search engine returns a list of ranked pages based on the relevance to the query. However, two major concerns are associated with the query-based Web search. The first problem is low precision, which is caused by a lot of irrelevant pages returned by the search engine. The second problem is low recall, which is due to the lack of capability of indexing all Web pages available on the Internet.

2. Finding needed information: 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 [12, 13].

3. Learning useful knowledge: With traditional Web search service, query results relevant to query input are returned to Web users in a ranked list of pages. In some cases, we are interested in not only browsing the returned collection of Web pages, but also extracting potentially useful knowledge out of them.

4. Recommendation/personalization of information: To improve the Internet service quality and increase the user click rate on a specific website, 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 [12,14].

There are three kinds of information that have to be dealt with when any user is accessing any web site [3].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 usage mining, web structure mining and web content mining [1,2]. Web content mining is the process of extracting knowledge from the content of documents. Web structure mining is the process of inferring knowledge from links between references and referents in the Web .Finally, Web usage mining, also known as Web Log Mining, is the process of extracting interesting patterns in web access logs .The main goal of Web usage mining is finding out what users are looking for on the Internet [3].

In this paper we are proposing an Approach to hybridize web content, web structure & web usage mining for Enhancing Web Search Engine Results Delivery. For Web Content mining a term-based weighted technique will be used to combine with Weighted Page Rank algorithm of Web structure mining & Web server log files to discover useful information.

The rest of the paper is organized as follows. A brief background review of web mining is presented in the next section. Section 3 presents the working of Web Search Engine. Section 4 presents the Literature Review. The proposed frame work of our approach is described in Section 5. Section 6 summarizes the conclusion.

II. WEB MINING

Oren Etzioni was the person who coined the term Web Mining first time [11].

Web mining is the Data Mining technique that automatically discovers or extracts the information from web. It consists of following tasks [3]:

1. *Resource finding:* It involves the task of retrieving intended web documents. It is the process by which we extract the data either from online or offline text resources available on web.

2. *Information selection and pre-processing:* This process transforms the original retrieved data into information. The transformation could be renewal of stop words, stemming or it may be aimed for obtaining the desired representation

3. *Generalization:* It automatically discovers general patterns at individual web sites as well as across multiple sites. Data Mining techniques and machine learning are used in generalization

4. *Analysis:* It involves the validation and interpretation of the mined patterns.

A. Web Mining Categories

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

1. **Web Content Mining**

Web content mining is the process of extracting useful information from the contents of web documents. Content data is the collection of facts a web page is designed to contain. It may consist of text, images, audio, video, or structured records such as lists and tables. It is related to text mining because much of the web contents are texts and Web data are mainly semi-structured and/or unstructured.

2. Web Structure Mining

Web Structure Mining (WSM) tries to discover the link structure of the hyperlinks in inter documents level. So web structure mining categorizes the web pages on the basis of the hyperlink and finds the similarity and relationship of information between different Web sites [16,17]. WSM explores the structure of the link inside the hyperlink between different documents and classify the pages of web. The number of out links i.e. links from a page and the number of in links i.e. links to a page are very important parameter in the area of web mining. 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. This can be further divided into two kinds based on the kind of structure information used.

Hyperlinks

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. A hyperlink that connects to a different part of the same page is called an *intra-document hyperlink*, and a hyperlink that connects two different pages is called an *inter-document hyperlink*.

Document Structure

In addition, the content within a Web page can also be organized in a tree structured format, based on the various HTML and XML tags within the page. The various contents of Web structure mining are

- Links Structure Mining

- Internal Structure Mining

- URL Mining

Links Structure: It consists of Link-based Classification, Link-based Cluster Analysis, Link Type, Link Strength and Link Cardinality.

Internal Structure Mining: It can provide information about page ranking or authoritativeness and enhance search results through filtering.

URL Mining: It gives a hyperlink which is a structural unit that connects a web page to different location, either within the same web page (intra_document hyperlink) or to a different web page (inter_document) hyperlink.

3. Web Usage Mining

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

Web usage mining itself can be classified further depending on the kind of usage data considered [8,10]:

Web Server Data

User logs are collected by the web server and typically include IP address, page reference and access time.

Application Server Data

Commercial application servers have significant features and ability to track various kinds of business events and log them in application server logs.

Application Level Data

New kinds of events can be defined in an application, and logging can be turned on for them generating histories of these events.

III. WORKING OF WEB SEARCH ENGINE

The World Wide Web consists billions of web pages and huge amount of information is available within the pages [5]. To retrieve required information from World Wide Web, search engines perform number of tasks based on their respective architecture. 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, Web Search Engine is a tool enabling document search with respect to specified keywords in the web and returns a list of documents where the keywords were found. A web search engine is a software system that is designed to search for information on the World Wide Web. The search results are generally presented in a line of results often referred to as Search Engine Result Pages (SERPs) [7].

There are 3 important components of search engine, Crawler, Indexer and Ranking mechanism. The crawler is also called as a robot or spider that traverses the web and downloads the web pages. The downloaded pages are sent to an indexing module that parses the web pages and builds the index based on the keywords in those pages. An index is generally maintained using the keywords [9]. 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 key features helping to retrieve a good representation of documents/web pages are as follows.

- **Term frequency**: How frequently a query term appears in a document is one of the most obvious ways of determining a document's relevance to a query.

- **Location of terms:** Many search engines give preference to words found in the title or lead paragraph or in the metadata of a document. Terms occurring in the title of a document or page that match a query term are frequently weighted more heavily than terms occurring in the body of the document. Similarly, query terms occurring in section headings or the first paragraph of a document may be more likely to be relevant.

- **Link analysis:** Link analysis is based on how well-connected each page is, as defined by Hubs and Authorities, where Hub documents link to large numbers of other pages (out-links), and

Authority documents are those referred to by many other pages, or have a high number of "in-links".

• **Popularity:** Google and several other search engines add popularity to link analysis to help determine the relevance or value of pages. Popularity utilizes data on the frequency with which a page is chosen by all users as a means of predicting relevance.

• **Date of Publication:** Some search engines assume that the more recent the information is, the more likely that it will be useful or relevant to the user. The engines therefore present results beginning with the most recent to the less current.

• **Length:** It is a factor used to compute the relative merit of similar pages. So, in a choice between two documents both containing the same query terms, the document that contains a proportionately higher occurrence of the term relative to the length of the document is assumed more likely to be relevant.

• **Proximity of query terms:** When the terms in a query occur near to each other within a document; it is more likely that the document is relevant to the query than if the terms occur at greater distance.

• **Proper nouns:** sometimes have higher weights, since so many searches are performed on people, places, or things.

IV. LITERATURE REVIEW

World Wide Web has become a powerful platform to store, disseminate 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 and H. Blockeel [3] presents 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.

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.

Brin and Page [16] developed Page Rank algorithm during their PhD at Stanford University based on the citation analysis. Page Rank algorithm is used by the famous search engine, Google. They applied the citation analysis in Web search by treating the incoming links as citations to the Web pages.

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.

Methods of organizing search results based on text categorization are studied in [18]. In this work, a text classifier is trained using a Web directory and search results are then classified into the predefined categories.

Clustering search results [19] is an effective way to organize search results, which allows a user to navigate into relevant documents quickly. The general idea is to perform clustering on a set of top- ranked search results to partition the results into natural clusters.

Wang and Zhai [20] clustered queries and learned aspects of these similar queries, which solves the problem in part.

H-J Zeng et.al [21] proposed a query based method to cluster search results. For a given query, the rank list of documents return by a certain Web search engine.

H. Chen and S. 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 which is 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.

V. FRAME WORK OF OUR APPROACH

To increase an accuracy factor and relevancy for information retrieval, the emphasis is given on the concept of Web Content, Web Structure and Web Usage mining techniques.

In this proposed work, the hybrid approach is used with the combination of text, hyperlinks and web log technique.

a) For Web Content mining a term-based weighted technique will be used to mine the web contents which will improve the performance of search engine results in terms of precision, recall and F-measure notations.

b) In web structure mining, Weighted Page Rank algorithm (WPR) takes into account for the importance of both the inlinks and the outlinks of the pages which distributes rank scores based on the popularity of the pages.

c) Currently the available search engines often give a long list of results, much of which are not always relevant to the user's requirement. Therefore Web Logs are important for information repositories, which will record the user activities on search results. This approach is based 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 result pages. This proposed system proves to be efficient as the pages desired by the user will be on the top priority in the result list and also optimize the query performance in terms of query results. The proposed methodology can be incorporated within a web based search engine to provide better performance.

VI. 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 research work will implement an efficient approach 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.

ACKNOWLEDGMENT

Ranjit Keole is thankful to Dr. P. P. Karde, Head of Department, Dept. of Information Technology, Govt. Women's Residential Polytechnic, Yavatmal, for his constant support and helping out with the preparation of this paper. He is also thankful to the Dr. V. M. Thakare, Professor & Head of Department, P. G. Department of Computer science, Sant Gadge Baba Amravati University, Amravati for being a constant source of inspiration.

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, Signed Approach 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. I. Mele, " Web Usage Mining for Enhancing Search –Result Delivery and Helping Users to Find Interesting Web Content," ACM SIGIR Conf.Research &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. 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 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.

11. O. Zamir and O. Etzioni, " Web Document Clustering: A Feasibility demonstration," ACM (SIGIR, 99) , pp. 46-54.

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 Con.on Data Storage and Data Engineering (DSDE), pp.187 – 191,*2010*.*

14. R. Kosala and H.Blockeel, "Web Mining Research: A Survey", SIGKDD Explorations ACM SIGKDD, July 2000.

15. WangBin and LiuZhijing , "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.

17. Wenpu Xing and Ali Ghorbani," Weighted PageRank Algorithm", IEEE, 2004.

18. Mele "Web Usage Mining for Enhancing Search-Result Delivery and Helping Users to Find Interesting Web Content",ACM SIGKDD, ,2013.

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. 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. H. Chen and S. 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. R.Bhushan and R. Nath, “Recommendation of Optimized Web Pages to Users Using Web Log Mining Techniques”, Published by the IEEE Computer Society, IEEE 2012.