Engineering Universe for Scientific Research and Management
(International Journal)
Impact Factor: 3.7
ISSN(ONLINE) : 2319-3069
Vol. 6 Issue 7
July 2014

# A Review – Mining Web Logs to Improve Website Organization

Bhavna Thakre[1] and Er. Kuntal Barua[2]
Computer Science & Engineering, RGPV, LNCT Indore, M.P,India[1]
Computer Science & Engineering, RGPV, LNCT Indore, M.P,India[2]
Bhavnathakre.it@gmail.com[1], kuntal.barua@gmail.com[2]

## Abstract

Web site design is critical to the success of electronic commerce and digital government. Effective design requires appropriate evaluation methods and measurement metrics. We define Web site navigability as the extent to which a visitor can use a Web site's hyperlink structure to locate target contents successfully in an easy and efficient manner. In this research, we propose an algorithm to automatically find pages in a website whose location is different from where visitors expect to find them. To find the required information accurately two or more web log files are merged. We present an algorithm for discovering such expected locations that can handle page caching by the browser. Expected locations with a significant number of hits are then presented to the website administrator. We also present algorithms for selecting expected locations (for adding navigation links) to optimize the benefit to the website or the visitor. Create an graphical representation of particular web pages to reduce the searching complexity of user. It also focus on those pages which got minimum number of user's click because that pages also contain some important information regarding the users search. We also have proposed using a time Threshold to distinguish between the two activities.

## Keywords:

*Web site design, Web log mining, Web site* navigability, page caching

## Introduction

The research is based on web mining which is classified into three types Web content mining, web usage mining and web structure mining. Web Structure Mining is the process of Discovering structure Information from the Web. 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 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. Mining efforts here have focused on automatically extracting document object model (DOM) structures out of documents. The structure of a typical Web graph consists of Web pages as nodes, and hyperlinks as edges connecting related pages.

In general, it is hard to organize a website such that pages are located where visitors expect to find them. This problem occurs across all kinds of websites, including B2C shops, B2B marketplaces, corporate websites and content websites. We propose an algorithm to solve this problem by discovering all pages in a website whose location is different from the location where visitors expect to find them. Expected locations with a significant number of hits are presented to the website administrator for adding navigation links from the expected location to the target page.

To find the required information accurately two or more web log files are merged. We present an algorithm for discovering such expected locations that can handle page caching by the browser. Expected locations with a significant number of hits are then presented to the website administrator

We also present algorithms for selecting the set of navigation links to optimize the benefit to the website or the visitor, taking into account that users might try multiple expected locations for a target page.We have proposed using a time threshold to distinguish between the two activities

## RELATED WORK

In this section we are going to discuss about the different types of web logs mining. Here we will study about the research work of different authors, how they use the

protocol to solve the problems and what are their future works.

**Hao Wu, Guoliang & Lighu Zhou**: This study is focus on to improve the compression ratio of inverted index using documents reordering. The documents ID's are reassigned so that similar documents are close to each other. It present an index called GINI Index which support key word search and encodes the documents Id's into intervals list and dramatically reduce the size of inverted index.

**Jingtian Jiang, Xinying Song, Nenghai Yu and Chin Yew Lin:** With this research the focus crawling problem to URL type reorganization is reduced and shows how to leverage implicit navigational path obtained by forums. For convenience of searching paths it is classified into three types' entry page, index page and thread page.

**Faustina Johnson & Santosh Kumar:** Proposed an algorithm for web content data. It considers the web search as general problem of function optimization. Using the fact the web graph is considered in which nodes are web pages and edges are links between web pages. Search Space optimization problem is a set of web pages. Fitness function is applied on the sets. Genetic algorithm is used to get top links.

**Charushila Kadu, Praveen Bhanodia, Pritesh Jain**: It involves conversion of unstructured into structured text input. Database identify the pattern and trends from the other structured database and finally extract useful information from text for this evaluation matrix is used which shows he overall performance of data and perform topic tracking with respect to a costant threshold to gather the related topic. Summarization removes the unnecessary information and reduces the length of document.

**Sunita Sharma and Ashu Bansal:** This presents the preprocessing step is used to give a reliable input for web mining task preprocessing is used to extract the important patterns from weblogs. For this statics analysis and data mining is applied on web logs. Data cleaning removes records of graphics, video and format information for that server log files are cleaned the data then user abd session identification is performed and the path and standardized data is given as a result.

**Aditi Shrivastava and Nitin Shrivastava**: To enhance the quality of website depends on the web access log file. When web user interact with a site data recording their behavior is stored in web server logs. Four types of server logs are used to test user's behavior common log, agent log error and reference log. It also gives information about pattern extract from web logs by this can shorten the pages which are not in the user pattern.

**Gajendra Singh and Priyanka Dixit:** The main objective of the proposed concept is to reduce total number of elements in each candidate set without any repeating the step which allow changes in large log record sets. It works on the achieved real time updates. The page which has minimum number of clicks rate will also finds association rule.

**V Shanmuga, Priya, and S Sakthivel:** The main purpose of the research to extract required pattern by removing noise that is present in web documents. A new architecture is created to manage the web data and during web content extraction method the unnecessary information is removed from the web information and creates the patterns for that data.

**Ramya C, Shreedhar K S and Kavitha G:** It proposed the possibilities of analyzing jointly multiple web server logs. It employs effective heuristics for detecting and eliminating web requests. It describes a general methodology for preprocessing the raw web logs into a structured form.

**Sanjeev Dhawan and Swati Goel:** The web logs file records information of each click by the user log files. Usually contain noisy and irrelevant data. Preprocessing is done to remove unnecessary data from log files. User and session identification is also done as a preprocessing step which includes identification user as a session pattern discovery technique such as association rule mining clustering & classification are applied on the reduced log files.

**Shaily langhnoja and Mehul P Baro**t: Basic association rule mining is may have drawback of generation of irrelevant rules. Generation of too many rules leading to contradiction prediction resulting in reduction in accuracy. So minimum support and confidence parameters can be set to eliminate false discoveries. Clustering frequent access pattern reduce data set for association rule mining and improve result accuracy and producing result of pattern discovery of web usage mining effective.

## LITERATURE EXTRACTION

Web site navigability as the extent to which a visitor can use a Web site's hyperlink structure to locate target

contents successfully in an easy and efficient manner out of all the above research works some authors like Faustina Johnson & Santosh Kumar use a graph to show web site structure. Ramya C, Shreedhar K S and Kavitha G give the method to jointly use multiple web log files. Hao Wu, Guoliang & Lighu Zhou find the desired web page by using Gini index in web log file. Sunita Sharma and Ashu Bansal give the concept of extracting useful information from the web logs and then by Gajendra Singh and Priyanka Dixit research reduces the element from candidate set.

## PROBLEM DEFINITION

Considering all the research the main problem is to reduce the unnecessary data from the weblog files and extract the important information. The overall focus is only on user behavior and users visiting pages rest of the pages are not taken into consideration. Web pages searching in not perform properly so that some important and necessary data is not received by user. Lots of burden is on index page so that the link complexity on index page is much more. Links between the web pages is much more Complex in structure.

## PROBLEM SOLUTION

There is many research is done by authors in field of web mining but the data is reduced so that the time consumption in searching is minimized and useful data is obtained. Server log files are merged to obtain much more related information regarding users click. Provide navigational links to users for particular link. The same and related keywords are taken into same interval of index. It helps in reconstructing or redesigning of website. The graphical representation of web pages is created to reduce structure complexity.

## CONCLUSION

The current research examines Web site navigability with a particular focus on the structural aspect of Web site design. The research is based on extracting the web logs data in efficient way and to reduce the linking complexity of web pages. Redesigning and reconstruction of web pages are making easier with this methodology. The graphical representation of web pages is created to reduce structure complexity. To increase the user click's related information two or more server log files are merged and then extract the information from that.

## REFERENCES-

[1]. Hao Wu, Guoliang & Lighu Zhou "Ginix: Generalized Inverted Index for keyword search" Year 2013, IEEE transaction on knowledge and Data mining Vol 8.
[2] Jingtian Jiang, Xinying Song, Nenghai Yu and Chin Yew Lin "FOCUS: Learning to Crawl web Forum" June 2013 IEEE transaction on Knowledge and web engineering Vol 25.
[3]. Charushila Kadu, Praveen Bhanodia, Pritesh Jain "Review of ontological based pattern mining techniques" 5 may 2013 International Journal of scientific and engineering research Vol 4.
[4]. Faustina Johnson & Santosh Kumar "Web content mining using genetic algorithm" 2013 Springer- Verlag Berlin Heidelberg.
[5]. Sunita Sharma and Ashu Bansal "Web Usage mining: Preprocessing of web log files" 4 april 2013 International Journal of advanced research in computer engineering and technology, Vol 2.
[6]. Aditi Shrivastava and Nitin Shrivastava "Extracting knowledge from user access logs" 4 april 2012 International Joural of Scientific and research publication, Vol 2.
[7]. Gajendra Singh and Priyanka Dixit "A new algorithm for web log mining" march 2014 International Journal of computer application Vol 90-No 17.
[8]. V Shanmuga, Priya, S Sakthivel "Implementation of web personalization using web mining technology" june 2013 International Journal of computer science and mobile computing Vol. 2 Issue 6.
[9]. Ramya C, Shreedhar K S and Kavitha G "Preprocessing: A Prerequisite for discovering patterns in WUM Process" march 2013 International Journal of information and electronics engineering Vol. 3 No 2.
[10]. Sanjeev Dhawan and Swati Goel "Web usage mining: Finding usage pattern from web logs" 2013 American International Journal of Research in Science, Technology, Engineering & Mathematics.
[11]. Shaily langhnoja and Mehul P Barot "Web usage mining using association rule on clustered data for pattern discovery" june 2013 International Journal of data mining techniques and application.
[12]. Gopal Pandey, Swati Patel, Vidhu Singhal and Akshay Kansara "A Process oriented perceptron of personalization technique in web mining" January 2013 International Journal of science and morden engineering Vol 1, Issue-2.
[13]. Marjan Eshaghi, S. Z. Gawali "Web usage mining based on complex structure of XML for web ID'S" April 2013 International journal of innovative technology and exploring Engineering Vol 2, Issue 5.

[14]. Suresh Shirgave and Prakash Kulkarni "semantically enriched web usage mining for predicting user future movements" October 2013 International Journal of Web & Semantic Technology (IJWesT) Vol.4, No.4,
[15]. Abdelhakim Herrouz, Chabane Khentout Mahieddine Djoudi "Overview of Web Content Mining Tools" 2013 The International Journal of Engineering and Science vol 2.
[16]. Abdelhakim Herrouz, Chabane Khentout Mahieddine Djoudi "Overview Of Web Content Mining Tools" 06 June 2013 The International Journal Of Engineering And Science, Vol 2 Issue 6
[17]. Ketki Muzumdar, Ravi Mante, Prashant Chatur "Neural Network Approach for Web Usage Mining" May 2013, International Journal of Recent Technology and Engineering, Vol 2 Issue 2.

**First Author:** Bhavna Thakre, BE 2010 MTech perusing, Student at Laxmi Narayan College Of Technology, Indore, paper on web log mining published in IJCEM Journal on weblog mining

**Second Author:** Er. Kuntal Barua, BE, MTech, Assistant Professor at Laxmi Narayan College of Technology, Indore