Techniques of Visualization of Web Navigation System

Anshu Srivastava

anshu_qrat@yahoo.co.in

Anurag Banoudha

iisc.anurag@gmail.com

Abstract- **Being the largest source of knowledge and information Internet has highly affected the global life. Although lots of techniques has been developed but huge amount of fact still to be uncovered in this field and lots of problems to be solved. When a web user visits so many pages after some times he forgets his way and distracts from his path. We produced a model for visualization of web usage data and web browser history data. We can find out various privious visited web page links and the link of home page or root page by using a path tracking algorithm. Finally we can visualize the website structure and visited web path with the help of visited web page link. This work is in special reference to Smart City users**

Keywords- WWW, Web Usage Data, Web History Data, Web Browsers, Web Site Structure, Web Path Tracking Algorithm.

1. Introduction

It is the time of internet. World Wide Web has linked all over the word together. It has changed the entire global system by changing the way of busines, education and research system. It has become the largest information source in the planet. The internet has wrapped up the whole world. Lots of log data are available on the web. Lots of work have been done in web usage mining to understand the web structure and to collect some useful information from web log files. It is known that a single picture or image can explain hundred words . So the hidden information on the web should be extracted and visualize to better understand the users activity. Sometimes a user distracted from his path while openning web pages and got bore and confused about his location on weblinks, where he is actually and on which web page he is visiting. Finally he realizes that he lost his way and the result is time lost . Visualization of web usage data may become the solution of this problem. Many visualization tools and techniques are available to visualize the web information. Web mining process starts from log data recording and ends on visualization.

A. Web Mining:

Web mining is the application of data mining technique to discover and analize useful information from the web. Oren Etzioni (1996) [1] used the term web mining to denote the use of data mining technjques to descover pattetns, web services and documents and informations. Web mining is the combination of two fields: data mining and world wide web. Web mining can be categorized into three areas: web content mining, web structure mining, and web usage mining.

B. Web Content Mining:

It is the discovery or retrieval of useful information from web pages having contents like text, images, audios, and videos . Web content mining is a method of data mining techniques for relational databases. The web documents usually contain many types data such as text, images, audio, video, meta data and hyperlinks. Some of them are semi structured such as HTML documents or a more structured data like the data in the tables or database generated HTML pages, but most of the data is unstructured text data. Mostly, web data residing in web documents are unstructured.

C. Web Structure Mining:

It is a technique to analyze the link structure of web sites by using graph theory. It usually involves the analysis of in-links and out-links and has been used for search engine results ranking and other web applications. Web structure mining can be divided into two types: -
(1). Extracting patterns from hyperlinks in the web where hyperlink connects the web page to a different location and (2). Analyze the tree-like structure of web page structures to describe, HTML or XML tag usage.
R. Cooley, B. Mobasher and J. Srivastava (1998) [2] proposed a system WEBMINER to structure a website and to analyze user access patterns. They also differentiated the web content mining from two points of views: Information Retrieval and Database.

D. Web Usage Mining:

It is the application of data mining technique to discover usage patterns from web data or usage logs to understand user’s requirement. Web usage mining contains those techniques that could predict users, behavior while they interact with the web because usage log data contains the users, identity and their browsing history. It also contains IP addresses, page references and access time of user. Web usage mining consists of three phases: preprocessing, pattern discovery, and pattern analysis. Preprocessing is an important phase because it takes maximum part in mining process. It includes the tasks of raw data cleaning, user identification, session identification and path completion and construction of transactions. Data cleaning is the task of removing unnecessary irrelevant records. User identification is the process of associating page references with same IP address with different users. Session identification is the process of breaking user's page references into user sessions Path completion is use to include or fill missing page references in a session. Construction of transactions is used to know the users interest and navigational behavior.

Above, we have discussed web mining with its areas, web content mining, structure mining, and web usage mining. We assume that visited web page history is also part of web usage data. Visualization of web usage history data woul d be helpful for users and beneficial in web usage mining. We will discuss more about this in our proposed methodology section. We will discuss some web mining and visualization works of our great researchers in literature survey part.

2. Literature Survey:

Since, web usage mining became the favorite area for researchers to discover interesting and frequent user navigation patterns from web server logs. Liu, H., and Keselj, V., (2007) [3] proposed a new approach to predict user's future request with analyzing and classifying navigation patterns. N. K. Tyagi, A.K. Solanki, and Manoj Wadhwa (2010) [4], analyzed the web server log files of smart sync software by web log expert program to get the information about errors and broken links which can be used by system administrators and web designers to website's effectiveness.

V.chitra, A. S. Davamani (2010) [5], described about the accomplishment of path completion, finding content path set and travel path set which shows users interest Author analyzed and implemented a data preprocessing treatment system for web usage mining and log data that contains data cleaning, user identification, path completion and transaction identification. If a user requests a specific page from server entries like gif, Jpeg, etc. are also downloaded which are not useful for analysis are eliminated. The records with failed status are also eliminated. Automated programs like web robots, spiders and crawlers are also removed from log files. Reference Length is the time taken by the user to view a particular page. It is computed by considering the byte transfer rate. Generally it is calculated by difference between access time of a record and the next record. Travel path transactions are constructed to know the navigational behavior of users. Auther used these preprocessing steps to give a reliable input for data mining tasks.

Sarah J. Waterson, Jason I. Hong, Jeffrey Heer, Tara Mathews (2003) [6] introduced a Web Quilt Visualization System for analyzing remote web usability click stream data gathered by WebQuilt proxy logger. Semantic zooming and filtering of click stream data is shown to be an effective method for exploring and probing usability data, allowing a designer to investigate the data for interesting issue within the context of the relevant web pages and tasks. WebQuilt is a proxy logger and a visualization system which can provide a useful usability information with combining the suitable testing analysis tools. Murat Ali Bayir, Ismail Hakki Toroslu(2009) [7] proposed a framework "SmartMiner" to create accurate user sessions and frequent navigation patterns in web usage mining. Since simple sessions are sequences of web pages requested from web server or visited in the web browsers based on time and navigation. Smart Miner sessions are set of paths traversed in web pages. They showed session creation as a new graph problem and used smart SRA algorithm to solve this problem efficiently.

A.H. Youssefi, D.J.Duke, M.J.Zaki (2004) [8] proposed a technique Visual Web Mining. They applied Data Mining techniques to large web data sets and used Information Visualization methods on the results. Their aim was to generate a combined visual graph structure of web pages and missing web usage log results. Jiyang Chen et al. (2004) [9] proposed a visualization tool to visualize web graphs with using web graph algebra to show intersting and hidden features of web data. Much taxonomy of information visualization techniques have been created using data centric point of view. Card and Mackinlay (1997) [10] had given a taxonomy in which visualization is devided in several subcategories: scientific visualization, GIS, multidimentional plots, multi dimentional tables, information landscapes and spaces, node and links, trees and text transforms. Online Library of Information Visualization Environment (OLIVE, 1999) [11] provided eight visual data types: Temporal, 1D, 2D, 3D, multi-D, Tree, Network and work space.

Ed H. Chi et al (1998) [12] also used processing operators with data types in taxonomy of visualization. Ed H. Chi (1999) [13] described Data Reference Model. In which he devided each technique into four stages of data, three types of data transformation and four types of stage operators. These four data stages are known as: value, analytical abstraction, visualization abstraction and view. Data moves from one stage to another in visualization data pipeline with using three types of operators known as date transformation, visualization and visual mapping transformation. Mihael Ankerst (2001) [14] used pixel oriented visualization techniques which is proposed by Daniel A. Kiem (1994) to map each attribute value of the data to a single colored pixel, representing the maximum amount of information. M. Ankerst, D. A. Kiem and H. P.Kriegel (1996) [15] described Cicle Segments Technique for visualizing large amounts of high dimensional data. Ed H. Chi (2002) [16] created a predictive visualization model called Information Scent; to decrease deficiencies in information accessibility and to find uncover patterns. His Scent Viz visualization system helps in analysis of large web usage data efficiently.

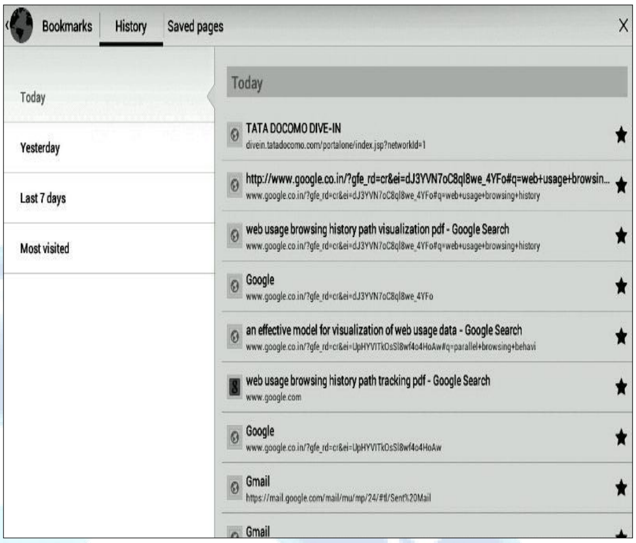
B. Zhou, et al. (2004) [17] proposed a Web Access Monitoring System for automatic discovery of visualization of user's temporal based web access behavior from client side logs. E. Herder, H. Weinreich (2005) [18] presented a web usage analysis tool "Navigation Visualizer " to select and match the data dynamically and to explore the graph based visualization interactively. It helps in tracing and understands user activities and provides means for preprocessing the complex usage data for statistical analysis. N. Labroche et al.(2008) [19] proposed a new tool for web usage mining and visualization which is based on the bio mimetic relational clustering algorithm "Leader Ant " to produce an efficient visualization of user's activity on website. R. S. Kasana et al.(2009) [20] described about the involvement of human factors in visualization process and the importance of human factors in creation of visualization tools. Since the discovery of web usage patterns would not be very useful without understanding them. Different tools and techniques are used for pattern analysis, visualization is one of them. Visualization is used to represent mined data graphically.

O. Hoeber et al.(2009) [21] proposed a method "BrowseLine” to guide users in re-finding web pages in their browsing histories. To represent data macro and micro both time levels two dimensional times line is used. A time line navigation view provides a zoomed out representation of the temporal features of browsing history. This also supports the visual recognition of patterns that match the user's re-collection of their browsing activities. Once an approximate area in th e timeline is identified, user can quickly jump to this location and evaluate the contents of the domain stacks as they seek to re-find a web page. A. Herrouz et al. (2013) [22] introduced the concept for web navigation with over viewing the several graphical navigation tools and technique. Web users face problems of locating themselves with respect to space and time during navigation through web. Due to lose interlinking between documents web users forget which page they visited and which page to visit next. Information searching became very difficult and time consuming due to immense growth of web.

3. Proposed methodology:

1. Our aim is to produce a model for visualization of web usage data. Since web usage mining is done to discover useful patterns in order to understand user's behavior and future interest. Here, we will try to combine the web usage data and web browser's history data. Web usage data can be gathered from different levels such as client level logs, server level logs, proxy server logs, cookies, and web browser's history at client side are collected. We have to find out the previous visited web page links and the link of home page or root page link by using a suitable path tracking tool or algorithms like Leader Ant Algorithm by N. Labroche (2007) [23] or path tracking algorithm proposed by Yan Li (2008) [24] And then we try to visualize them.
2. Since web usage data is main part of this research, we have to know that where usage data is stored and how can we visualize the structure of web usage data and visited web page link structure of web sites. We should know what are the benefits of this process? As we seen in the paper "education trails system", by C. Romero, et. al, 2008) [25], students visit web pages several times or visit any page, all information or files are stored in a usage log in (. dat) format or in any other format. We should know about the web page visit time and visiting frequency. Those pages which are missing they can be identified with the help of suitable path completion techniques.
3. Three main phases of web usage mining are very important that are: data preprocessing, knowledge extraction and analysis of extracted results. Data preprocessing contains some tasks: data cleaning, user identification, session identification, path completion, and transaction identification. After collection of web usage data from different sources like client logs, server logs, cookies and web browsers histories, we can create a proper link structure of the web pages and then visualize these structure so that user can better understand. This process would be useful to decrease the path distraction and time loss problem due to confusion [26].

Following figures are given for illustration:



TATA DOCOMO DIVE-IN
divein.tatadocomo.com/portalone/index.jsp?networkId=1

http://www.google.co.in/?gfe_rd=cr&ei=dJ3YVN7oC8ql8we_4YFo#q=web+usage+browsin...
www.google.co.in/?gfe_rd=cr&ei=dJ3YVN7oC8ql8we_4YFo#q=web+usage+browsing+history

web usage browsing history path visualization pdf - Google Search
www.google.co.in/?gfe_rd=cr&ei=dJ3YVN7oC8ql8we_4YFo#q=web+usage+browsing+history

Google
www.google.co.in/?gfe_rd=cr&ei=dJ3YVN7oC8ql8we_4YFo

an effective model for visualization of web usage data - Google Search
www.google.co.in/?gfe_rd=cr&ei=UpHYVITkOsSl8wf4o4HoAw#q=parallel+browsing+behavi

web usage browsing history path tracking pdf - Google Search
www.google.com

Google
www.google.co.in/?gfe_rd=cr&ei=UpHYVITkOsSl8wf4o4HoAw

Gmail
https://mail.google.com/mail/mu/mp/24/#tl/Sent%20Mail

Gmail

Figure 1: Web browser history

As when a user reads a book, He generally starts from index page or first page where topic name is initialized which is connected to every page of the book. Although, when we read any book we usually move one page to another one by one. Because these pages are mutually linked with two ways: First, by page numbers, which give us the information about page position and second by topic events which are mutually related. But we don't know whether each page of the book is connected to its start page page or not. Since each web page of a web site is connected through hyperlinks or buttons. When user sees any web page url stored in a web browser history, he does not know that which is the previous or next page of this page url. In some websites these facilities are available for example in Wikipedia, mostly word hyperlinks and image hyperlinks are available on its any web page. It may create some confusions for novel users to understand how to move on next or how to return home page [27].

The following figure shows that a web site also can be shown in the form of a matrix structure. When a coefficient is multiplied by any matrix, it moves to relate all elements of the matrix similarly the home page also can be connected to all pages of a website. And when we visualize the matrix structure of any website then we may get result.

There are so many different visited web page urls are stored in web browser's histories. Here the question is that can we get the home page of this visited url page, if yes by any means then we can easily visualize the matrix structure of the whole website. The point is that we can get the whole information from a single visited page url and visualized structure. This could be beneficial to improve the websites efficiently and for users for easy net surfing.

4. Conclusion:

In paper, we have given some overview on web mining, web content mining, web structure mining and web usage mining. We described about web usage data logs and web usage data history in web browsers. These descriptions are to given to point out the users problems during net surfing or visiting web pages. The main aim our work is to find out a way to make netsurfing easy for every individual user. Although great researchers provided lots of tools and techniques to improve the websites efficiently. With the help of their achievements we can get some beneficial results. We tried to propose an methodology to combine the web usage log mining with web usage data histories in a browser at client side. With the help of mining and visualizing the visited web path and structure of the website from web browser histories we can get the whole information about a website.

References:

[1]. Oren Etzioni: "The World Wide Web: Quagmire or gold mine". Communication ACM 39(11):65-68, 1996.

[2]. R. Cooley, B. Mobasher and J. Srivastava "Web Mining: Information and Pattern Discovery on the World Wide Web" in proceeding of the 9th IEEE International Conferece on tool with artificial intelligence (ICTAI' 97).

[3]. Liu H. and Keselj V., "Combined mining of server logs and web contents for classifying user navigation patterns and predicting user's future requests", Data and Knowledge Engineering, 2007, Vol 61, Issue 2, PP. 304-330.

[4]. N. K. Tyagi, A. K. Solanki and Manoj Wadhwa, "Analysis of server log by web usage mining for website improvement " International journal of computer science issue (IJCSI), Jul 2010, vol. 7 Issue 4, p17, 2010.

[5]. V. Chitra, A. S. Davamani, "An EfficientvPath Completion Technique for Web Log Mining", 2010 IEEE, International Conference on Computational Intelligence and Computing Research, ISBN: 9788183713627.

[6]. Sarah J. Waterson, Jason I. Hong, Jeffrey Heer, Tara Mathews "What Did They Do? Understanding clickstreams with the WebQuilt Visualization System ", May 2002.

[7]. Murat Ali Bayir, Ismail Hakki Toroslu, "SmartMiner: a new framework for mining large scale web usage data", 2009.

[8]. A.H. Youssefi, D. J. Duke, M. J. Zaki," Visual Web Mining", www2004, May 17-22, 2004, New York, NY, USA.

[9]. Jiyang Chen, Lisheng Sun, Osmar R. Zakiane, Randy Goebel, "Visualization and Discovering Web Navigational Patterns ", 7th International Workshop on the Web and Databases(WebDB 2004) June 17-18, 2004, Paris, France.

[10]. S. K. Card and J. D. Mackinlay, "The structure of the Information Visualization Design Space", proceeding of IEEE Symposium on Information Visualization (InfoVis '97), Phoenix, Arizona, 92-99, Color Plate 125, 1997.

[11]. On-Line Library of Information Visualization Environments (OLIVE, 1999). http://octal.umd.edu/

[12]. Ed H. Chi and J. T Riedl. "An Operator Interaction Framework for Visualization Systems", Symposium on Information Visualization (InfoViz ' 98), Research Triangle Park, North Carolina: 63-70, 1998.

[13]. Ed H. Chi, "A Framework for Visualization Spreadsheets", Ph.D. Thesis, University of Minnesota, March 1999.

[14]. M. Ankerst, "Visual Data Mining with Pixel-Oriented Visualization Techniques ", The Boeing Company, P.O. Box 3707 MC 7L-70, Seattle, WA 98124.

[15]. M. Ankerst, D. A. Kiem and H. P. Kriegel, "Circle Segments: A Technique for Visually Exploring Large Multidimensional Data Sets", Proc. Visualization '96, Hot Topic Session, San Francisco, CA, 1996.

[16]. Ed H. Chi, "Improving Web Usability Through Visualization", Palo Alto Research Center, IEEE, Internet Computing, March-April 2002. http://computer.org/internet/1089-7801/02.

[17]. B. Zhou, C. Hui, and Alvis C. M. Fong, "School of Computer Engineering, Nanyang Technological University, Singapore ".

[18]. E. Herder, H. Weinreich, "Interactive Web Usage Mining with the Navigation Visualizer ", CHI, 2005, April 2-7, 2004, Portland, Oregon, USA, ACM 1-59593-002-7/05/0004.

[19]. N. Labroche, Marie-Jeane Lesot, L. Yaffi, "A new web usage mining and visualization tool", Universite Pierre et Marie Curie-Paris 6, UMR 7606, Paris, F- 75005 France, Intelligent Learning Objects, 2008.

[20]. R. S. Kasana, Ratnesh Kumar Jain, Suresh Jain, "Visualization of Mined Pattern and its Human Aspects", IJCSIS, Vol. 4, No. 1 & 2, 2009.

[21]. O. Hoeber, Joshua Gorner,"BrowseLine: 2D Timeline Visualization of Web Browsing Histories." 2009.

[22]. Herrouz, C. Khentout, and Mahieddine Djoudi, "Overview of Visualization Tools for Web Browser History Data." 2013.

[23]. N. Labroche, Learning web users profiles with relational clustering algorithms. In Workshop on Intelligent Techniques for Web Personalization AAAI Conference, 2007.

[24]. Yan Li, Boqin FENG and Qinjiao MAO, "Research on Path Completion Technique in Web Usage Mining ", International Symposium on Computer Science and Computational Technology, IEEE, 2008.

[25]. C. Romero, S. Gutierrez, M. Freire and Sebastian Ventura," Mining and Visualizing Visited Trails in Web-Based Educational Systems", 2008.

[26]. Anshu Srivastva, et. al., "KDD based System Making Information Assessment Sustainable for Web Application" International Journal of Research and Development in Applied Science and Engineering, Volume 4, Issue 1, November 2013.

[27]. Anshu Srivastava et. al.., "An Approach for Personalization on Web Based Systems" International Journal of Research and Development in Applied Science and Engineering, Volume 5, Issue 1, March 2014.