ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 7, Issue 10, October 2018

Web Usage Mining in Ranking Web Pages

Shadab Irfan[1], Subhajit Ghosh[2]

Research Scholar (SCSE), Galgotias University, Greater Noida, Uttar Pradesh, India.[1]

Professor (SCSE), Galgotias University, Greater Noida, Uttar Pradesh, India[2]

ABSTRACT: With the rapid increase in the size of the web data new methods and techniques have been emerging which help in accessing the information easily. Ranking of the pages are being done which help in finding relevant information as per need of the user. In order to facilitate the ranking process various algorithms are being designed from time to time which help to rank the pages in order of priorities. The ranking algorithms follow different web mining techniques like structural links, content format and usage pattern in order to rank the web pages. Web Usage Mining can be quite helpful in the process of ranking as they consider the web log files which store the usage pattern of the users. It helps to track the visiting pages on the basis of query. The navigational behavior of the users can be captured which ease the process of retrieval. The paper focus on basic web mining techniques and ranking algorithms and throws light on the effect of web usage mining in the ranking process and its overall strategy.

KEYWORDS: Web Mining, Web Usage Mining, Ranking Algorithms.

I. INTRODUCTION

Due to the rapid pace in technology the data in the World Wide Web is increasing at a rapid pace. It is a consortium where data of different formats are pooled together so that the user can access their required information. The search engine is designed to cater the need of the users [1]. As the data is in enormous form so it is quite difficult for the user to extract the needed information easily in constraint time period which is in different forms like text, image, video and audio [2]. The process of data mining helps in managing the repository and is quite useful in finding out the required patterns as per user need [3]. Various methods and tools have been designed for web mining that not only help to categories the data but also help in easy retrieval. Apart from these various ranking algorithms are designed that overall ease the task of arranging the web pages according to the priorities based on user queries. These algorithms have been modified from time to time so that more refined web pages can be displayed. The techniques being used are based on content and structural links while little work has been done on web logs file that incorporate web usage mining. Web Usage Mining will be quite helpful in information retrieval process as they take into consideration the usage pattern of the user which also helps in increasing the relevancy of the web pages. The paper throws light on the various ranking algorithms and the various web mining techniques that are used in these ranking algorithms with special focus on web usage mining techniques.

II. RELATED WORK

Web Mining has been used in extracting the patterns which are relevant to the user. In this regard WCM, WSM and WUM played a key role in extracting and ranking documents. Various researchers' focuses on the importance of Web Usage Mining and the different algorithms being employed are reviewed properly. Parth Suthar [3] et al. focuses on the Web Usage Mining Technique where he described the various algorithm which are used in pattern discovery phase of WUM and also provide their detail view along with their pros and cons. Neeraj Kandpal [4] et al. presented an outline view of the various log files which are used in WUM along with their applications. Anmol Kaur [5] et al. studied different algorithms used in WUM and focus on the Apriori and k-means clustering algorithm.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 7, Issue 10, October 2018

III. MINING TECHNIQUE

Web Mining is normally a sub field of data mining that is applied to data by using some intelligent techniques in order to find out interesting patterns that are relevant for decision making tasks. It sometimes also encompasses natural language techniques and artificial intelligent methods on semi structured and unstructured data so that relevant information can be retrieved. Web Mining is broadly categorized into Web Content Mining (WCM), Web Structure Mining (WSM) and Web Usage Mining (WUM) as depicted in Fig 1. WCM deals with the content of the web page and try to extract useful information found within the page. WSM works on the structural links of the web pages and try to trace web pages that are source of information. By considering links among the web pages it can be treated as a directed graph which captures the important links among the pages. WUM works on the web log files which are used to store the browsing pattern of the user who visited the web pages. On the basis of the navigational pattern the behavior of the user can predicted which also help in knowing the needs and interest of the user [3, 6, 7].

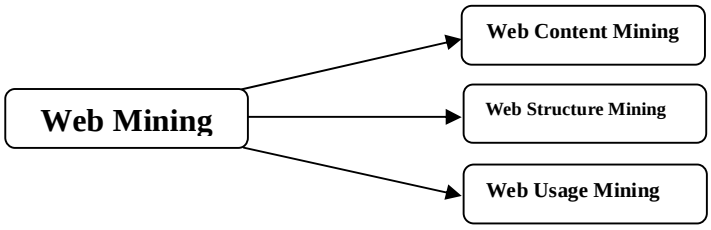

Fig 1: Web Mining Technique

During Web Usage Mining process the primary data source is log data which is stored in web log files. The web usage data that is collected from different sources can be categorized into-

- Usage Data
- Content Data
- Structure Data
- User Data

The Usage Data is considered as the primary data source that is collected from application and web servers and capture the fine navigational behavior of the users. The web objects that are collected during user event are basic level of pageview abstraction. When a user visits a site the sequence of pageview events are recorded. The Content Data include static pages and files which comprises of different objects and their relationship among each other. It may also include semantic or structural meta-data that is embedded within the site. The Structure Data focus on intra-page structure and represent the designer view for a particular site which can be represented in form of graph or tree. The HTML and XML can be represented in form of tree and the hyperlink structure can be captured by site map that is generated automatically. The User Data contain profile information about users. It comprises of demographic information about who visited the site regularly, ratings, visit histories and user interest information. [8]

IV. WEB USAGE DATA

During Web Usage Mining process the primary data source is log data which is stored in web log files. The web usage data that is collected from different sources can be categorized into-

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 7, Issue 10, October 2018

- Usage Data
- Content Data
- Structure Data
- User Data

The Usage Data is considered as the primary data source that is collected from application and web servers and capture the fine navigational behavior of the users. The web objects that are collected during user event are basic level of pageview abstraction. When a user visits a site the sequence of pageview events are recorded. The Content Data include static pages and files which comprises of different objects and their relationship among each other. It may also include semantic or structural meta-data that is embedded within the site. The Structure Data focus on intra-page structure and represent the designer view for a particular site which can be represented in form of graph or tree. The HTML and XML can be represented in form of tree and the hyperlink structure can be captured by site map that is generated automatically. The User Data contain profile information about users. It comprises of demographic information about who visited the site regularly, ratings, visit histories and user interest information [8].

V. WEB USAGE MINING TECHNIQUE

Web Usage Mining focuses on discovering and analysis of interesting patterns by capturing the behavior patterns of the users. The web usage data consists of web server logs, browser logs, proxy logs, cookies etc. It helps in analyzing the browsing patterns of the users by applying various data mining techniques like clustering, classification, association rules etc. The Web Log Files are the primary data source for WUM. The Web Usage Mining comprises of three inter dependent stages as depicted in Fig. 2. It describes the Data Collection, Pattern Discovery and Pattern Analysis stages.

- Data Collection and Pre Processing.
- Pattern Discovery.
- Pattern Analysis.

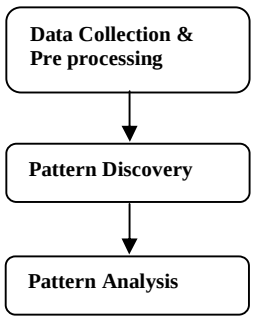

Fig 1: Web Usage Mining Technique

Data preprocessing involves different steps that helps in cleaning the data and identifying the user who visit the site along with the sequence of action that user performs .The usage preprocessing is regarded as most difficult as here the data is incomplete and help to trace user and session. The content preprocessing phase works on converting the files into a form which can be easily processed by the web usage mining process and employs technique like clustering and classification. The structure preprocessing employs hypertext links for site structure.
Data preparation is one of the important steps which require thorough analysis of the data by using heuristic techniques and is a time consuming. First the data that is collected during data fusion and cleaning process from the web log files

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 7, Issue 10, October 2018

need preprocessing. The items which are not required for analysis are removed in the cleaning process. It is required to mine the knowledge efficiently. After this in pageview and user identification steps the intra structure of the web page and its corresponding attributes are recognized and multiple sessions of the users are identified who visited the site. The sessionization helps to segment the user activity and record the sequence of actions which are performed by the users. The subsequence of user sessions is tracked in the episode identification which is the final step of preprocessing stage. After this the data from different sources are integrated which provide an effective framework for pattern discovery.

The second phase of Web Usage Mining is important as it helps to apply various techniques of data mining that unveils the hidden information and help in analyzing the patterns. During Pattern Discovery phase different forms of machine learning and statistical operations are undergone which help to obtain hidden patterns and thus reflect the behavior of the ongoing users, web resources and different form of summary statistics.

During association rule mining the hidden patterns are discovered and frequently used patterns are revealed. It tries to establish the relationship between the pages that have been visited. The clustering helps to group similar pages and find out users who have a common behavior. In usage domain two clusters are considered- user clusters and page clusters. Usage based clustering help in forming clusters for users who have same interest and can be mapped using stochastic methods and k-means algorithms. On the other hand the clustering of pages having similar content can be grouped together that will help in finding the search pattern of the users.

The sequential pattern phase helps to analyze the behavior of the user, in what order the web pages are visited which help in tracking the interest of the users. The psychology of the user can be tracked by following the navigation of the web pages. Statistical learning technique like Markov model can help in performing user navigational behavior and are quite helpful in predictive modeling which is based on sequence of events. Classification helps to classify the behavior of the user on basis of certain set of attributes and help in mapping the behavior of users using supervised learning techniques. It involves extracting selected features and then mapping them into different classes for further processing. A special kind of classification and prediction system referred as recommender system help in recommending products and services to the users on the basis of user interest and rating.

The motivation of the third phase Pattern Analysis lies in removing unwanted patterns of previous stages and the final patterns which are discovered are analyzed and validated for further operations. This phase help in discovering the patterns which overall can be employed in different visualization and report generation tools. Graphic patterns can help in highlighting the data trends and various patterns. The procedure helps to find out the model that follows a particular pattern. Knowledge querying and OLAP techniques are used for the task.

VI. AREAS OF WEB USAGE MINING

WUM helps in predicting future trends and behavior of the users which help in proactive decision making tasks. By analyzing the flow of data in a particular direction the shopping sites can give offers on certain range of products and at the same time the banks can give credits to its cardholders. It helps the organizations in various ways-

- Help in determining life-time value of its clients.
- Used for finding the cross marketing strategies among the organizations.
- For web applications help to optimize the functionality.
- Help in providing personalized content to the visitors [8].

Various data modeling techniques like ANN, Genetic algorithms, Decision Trees, classification, rule induction and data visualization can help in analyzing data and predicting patterns which can be converted into knowledge by using various graphical patterns. It helps in navigating the activities of the user and overall plays a key role in web designing and management.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 7, Issue 10, October 2018

VII. WEB RANKING ALGORITHM

Web mining techniques are employed on the search engines that help in retrieving the desired information as per need of the user. Various ranking algorithms are being proposed from time to time for the search engine which helps in efficiently extracting the information on the basis of query entered by the user. The ranking algorithms play a major role in arranging the web pages in order of relevance and help the user in finding the needed information in minimum amount of time. Retrieving quality pages is one of the main tasks of ranking which uses different metrics to check the quality of the pages on basis of user query [9]. These ranking algorithms are based on web content, structural links and web log files for extracting the needed information. Sometimes combinations of more than one parameter are merged together for efficient retrieval [2, 10].

The Page Rank, Weighted Page Rank, Distance Rank and Relation Based Algorithm are all based on the structural links among the web pages [11,12,13,14] . The Eigen Rumor, Tag Rank and Query Dependent Ranking Algorithm are based on the inner content of the web pages for ranking [15, 16, 17]. Hypertext Induced Topics Search (HITS) and Weighted Link Rank Algorithm incorporate both the content and structural links among the web pages while the Time Rank Algorithm follows the usage mining technique to arrange the web pages [18, 19, 20, 21]. All these algorithms focus on arranging the web pages by following the link structure of the pages or the content within the pages or by assigning weight on the basis of popularity of the pages. Some algorithms works for blog only while other considers the distance between the pages for calculating rank other work on visited time factor for ranking the pages. Overall all these algorithms help the user in finding the relevant web pages as per their needs.

VIII. IMPORTANCE OF WEB USAGE MINING IN RANKING

Most of the ranking algorithms focus on the content and structural links of the web pages to extract information. The significance of web usage mining is normally ignored in most of the ranking algorithms. The server log files serve as a great source of information as they carry the detailed information of the user navigational path. The Web Usage Mining considers the log data that is stored at the client side, server side and proxy side to enhance the working of the ranking algorithms. Normally data stored at client site in cookies are not authentic and are not considered for processing so log data stored at server side which comprise of various information like IP address, page requested, time stamp, navigational pages, bytes served, session information etc. are considered for processing [22].

The Web Usage Mining perform the necessary steps like preprocessing, pattern discovery and pattern analysis on the log data which help in identifying the information about particular users who visit the site and their area of interest. By analyzing the data it can be interpreted that certain pages are of interest to a particular group of user or certain items are favored by a group of user. Such types of activities help in increasing the business and enhancing the structure of the sites. By applying association rules, classification and clustering on the web pages the most frequent visited and associated pages can be traced and grouped together. These process can overall help in improving the ranking of the web pages so that the needed information can be traced easily. The Recommender System can be employed which help the lame user in recommending certain pages of interest on the basis of search query. By analyzing the navigational behavior of the user and pages of interest more similar pages can be recommended which help in the searching process

IX. CONCLUSION

The paper gives a glimpse of various Web Mining Techniques with special focus on Web Usage Mining. The Usage Mining has emerged as a new tool which can help in ranking of the documents. The ranking algorithms employ various techniques to retrieve relevant pages and in this perspective Web Usage Mining can play an important role. The Usage patterns can help in capturing the navigational pattern of the user which can overall help in finding similar pages as per need of the user thereby reducing time and complexity. The technique can be useful by employing intelligent methods and tools and overall ease the process of information retrieval from the web pages.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 7, Issue 10, October 2018

REFERENCES

[1] Ms. Shital C. Patil, Prof. R. R. Keole, "The Role of Web Content Mining and Web Usage Mining in Improving Search Result Delivery", International Journal of Computer Science and Mobile Computing, Vol.3 Issue.3, March- 2014.
[2] Shadab Irfan, Subhajit Ghosh," A Review on Different Ranking Algorithms", IEEE ICACCCN 2018.
[3] Parth Suthar, Prof. Bhavesh Oza, "A Survey of Web Usage Mining Techniques", International Journal of Computer Science and Information Technologies, Vol. 6 (6) , 2015.
[4] Neeraj Kandpal, Ripu Ranjan Sinha, M. S. Shekhawat,"A Survey On Web Usage Mining: Process, Application And Tools", Suresh Gyan Vihar University Journal of Engineering & Technology, Vol . 3, Issue 1, 2017, pp 19-25 ISSN: 2395-0196.
[5] Anmol Kaur and Dr. Raman Maini, "Analysis of Web Usage Mining techniques to predict the user behavior from Web Server Log Files",International Journal of Advanced Research in Computer Science, Volume 8, No. 5, May-June 2017.
[6] Shadab Irfan, Subhajit Ghosh, "Web Mining for Information Retrieval", International Journal of Engineering Science and Computing, Volume 8 Issue No.4 April 2018.
[7] Raymond Kosala & Hendrik Blockeel, "Web Mining Research: A Survey", SIGKDD Explorations, ACM 2000.
[8] Bing Liu," Web Data Exploring Hyperlinks, Contents, and Usage Data, Springer.
[9] Zakaria Suliman Zubi, "Ranking WebPages Using Web Structure Mining Concepts", Recent Advances in Telecommunications, Signals and Systems.
[10] N. V. Pardakhe, Prof. R. R. Keole, "Analysis of Various Web Page Ranking Algorithms in Web Structure Mining", International Journal of Advanced Research in Computer and Communication Engineering Vol.2, Issue 12, December 2013.
[11] Sergey Brin and Larry Page, "The anatomy of a Large-scale Hypertextual Web Search Engine", In Proceedings of the Seventh International World Wide Web Conference, 1998.
[12] Wenpu Xing and Ali Ghorbani, "Weighted PageRank Algorithm", Proceedings of the Second Annual Conference on Communication Networks and Services Research, 2004 IEEE.
[13] Ali Mohammad Zareh Bidoki, Nasser Yazdani, "DistanceRank: An intelligent ranking algorithm for web pages', Information Processing and Management, 2007 Elsevier.
[14] Fabrizio Lamberti, Andrea Sanna and Claudio Demartini , "A Relation-Based Page Rank Algorithm for. Semantic Web Search Engines", In IEEE Transaction of KDE, Vol. 21, No. 1, Jan 2009.
[15] Ko Fujimura,Takafumi Inoue and Masayuki Sugisaki, "The EigenRumor Algorithm for Ranking Blogs",WWW 2005.
[16] Shen Jie,Chen Chen,Zhang Hui,Sun Rong-Shuang,Zhu Yan and He Kun, "TagRank: A New Rank Algorithm for Webpage Based on Social Web" In proceedings of the International Conference on Computer Science and Information Technology, 2008.
[17] Lian-Wang Lee, Jung-Yi Jiang, ChunDer Wu, Shie-Jue Lee, "A Query-Dependent Ranking Approach for Search Engines", Second International Workshop on Computer Science and Engineering, Vol.1, PP. 259-263, 2009.
[18] Jon M. Kleinberg, "Authoritative Sources in a Hyperlinked Environment".
[19] H Jiang et al., "TIMERANK: A Method of Improving Ranking Scores by Visited Time", In proceedings of the Seventh International Conference Machine Learning and Cybernetics, Kunming, 12-15 July 2008.
[20] Ricardo BaezaYates and Emilio Davis, "Web Page Ranking using Link Attributes",May 17–22, 2004,ACM.
[21] Shadab Irfan, Subhajit Ghosh ,"Analysis & Challenges of Web Ranking Algorithms", ICCCA 2018.
[22] Ketan D. Patel,Satyen M. Parikh,"Preprocessing on Web Server Log Data for Web Usage Pattern Discovery", International Journal of Computer Applications,Volume 165 – No.10, May 2017.