ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

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

Website: www.ijircce.com

Vol. 5, Issue 11, November 2017

Survey on Data mining based prediction of User Behaviour through sessions

Tushar D. Kolhe[1], Nilesh Y. Choudhary[2]

P.G. Student, Department of Computer Engineering, GF'S Godavari Engineering College, Jalgaon, India[1]

Associate Professor, Department of Computer Engineering, GF'S Godavari Engineering College, Jalgaon, India[2]

ABSTRACT: Users are increasingly pursuing complex task-oriented goals on the web, such as making travel arrangements, managing finances, or planning purchases. To this end, they usually break down the tasks into a few codependent steps and issue multiple queries around these steps repeatedly over long periods of time. To better support users in their long-term information quests on the web, search engines keep track of their queries and clicks while searching online. In this paper, we study the problem of organizing a user's historical queries into groups in a dynamic and automated fashion. Automatically identifying query groups is helpful for a number of different search engine components and applications, such as query suggestions, result ranking, query alterations, sessionization, and collaborative search. In our approach, we go beyond approaches that rely on textual similarity or time thresholds, and we propose a more robust approach that leverages search query logs. We experimentally study the performance of different techniques, and showcase their potential, especially when combined together
.

KEYWORDS: Energy efficient algorithm; Manets; total transmission energy; maximum number of hops; network lifetime

I. INTRODUCTION

AS the size and richness of information on the web grows, so does the variety and the complexity of tasks that users try to accomplish online. Users are no longer content with issuing simple navigational queries. Various studies on query logs (e.g., Yahoo's and AltaVista's) reveal that only about 20 precent of queries are navigational. The rest are informational or transactional in nature. This is because users now pursue much broader informational and task oriented goals such as arranging for future travel, managing their finances, or planning their purchase decisions. However, the primary means of accessing information online is still through keyword queries to a search engine. A complex task such as travel arrangement has to be broken down into a number of co-dependent steps over a period of time. For instance, a user may first search on possible destinations, timeline, events, etc. After deciding when and where to go, the user may then search for the most suitable arrangements for air tickets, rental cars, lodging, meals, etc. Each step requires one or more queries, and each query results in one or more clicks on relevant pages.

One important step toward enabling services and features that can help users during their complex search quests online is the capability to identify and group related queries together. Recently, some of the major search engines have introduced a new "Search History" feature, which allows users to track their online searches by recording their queries and clicks. For example, a portion of a user's history as it is shown by the Bing search engine on February of 2010. This history includes a sequence of four queries displayed in reverse chronological order together with their corresponding clicks. In addition to viewing their search history, users can manipulate it by manually editing and organizing related queries and clicks into groups, or by sharing them with their friends. While these features are helpful, the manual efforts involved can be disruptive and will be untenable as the search history gets longer over time.

In fact, identifying groups of related queries has applications beyond helping the users to make sense and keep track of queries and clicks in their search history. First and foremost, query grouping allows the search engine to better understand a user's session and potentially tailor that user's search experience according to her needs. Once query groups have been identified, search engines can have a good representation of the search context behind the current query using queries and clicks in the corresponding query group. This will help to improve the quality of key

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

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

Website: www.ijircce.com

Vol. 5, Issue 11, November 2017

components of search engines such as query suggestions, result ranking, query alterations, sessionization, and collaborative search. For example, if a search engine knows that a current query "financial statement" belongs to a {"bank of america," "financial statement"} query group, it can boost the rank of the page that provides information about how to get a Bank of America statement instead of the Wikipedia article on "financial statement," or the pages related to financial statements from other banks.

Query grouping can also assist other users by promoting task-level collaborative search. For instance, given a set of query groups created by expert users, we can select the ones that are highly relevant to the current user's query activity and recommend them to her. Explicit collaborative search can also be performed by allowing users in a trusted community to find, share and merge relevant query groups to perform larger, long-term tasks on the web.

II. RELATED WORK

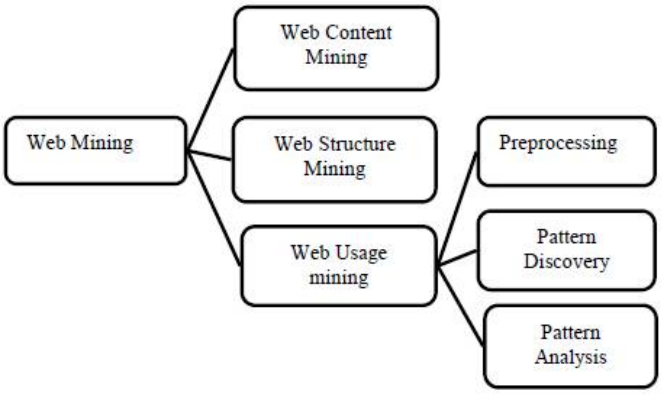

Fig 1: Web Mining Structure

Web Content mining [3] deals with discovery of useful information from unstructured, semi structured or structured contents of web documents. Text, images, audio, video comprised by unstructured document, semi structured data includes HTML documents and lists and tables represent structured documents. The main aim of web content mining is to act as tool to retrieve information easily and quickly. Web Content Mining works by organizing a group of documents into related categories which helps web search engine to ex-tract information more quickly and efficiently. Web Structure Mining [6], [7] mines the information by utilizing the link structure of the web documents. It works on inter document level and discovers hyperlink structure. It helps in describing the similarities and relationships between sites. Web Usage Mining [3] is a data mining technique that mines the information by analyzing the log files that contains the user access patterns. Web Usage Mining mines the secondary data which is present in log files and derived from the interactions of the users with the web. Web usage Mining techniques are applied on the data present in web server logs, browser logs, cookies, user profiles, bookmarks, mouse clicks etc. This information is often gathered automatically access web log through the Web server.

2.1 web usage mining

Web Usage Mining concentrates on the techniques that could predict the navigational pattern of the user while the user interacts with the web. It is mainly divided into two categories, they are general access pattern tracking and customized usage tracking. In general access pattern tracking information is discovered by using the history of web page visited by user while in customized usage tracking mining is targeted on specific user. Mainly there are four types of data sources present in which usage data is recorded at different levels they are: client level collection, browser level collection, server level collection and proxy level collection.

Client Level collection: At this level data is gathered together by means of java scripts or java applets. This data shows the behaviour of a single user on single site. Client side data collection requires user participation for enabling java scripts or java applets. The advantage of data collection at client side is that it can capture all clicks including pressing of back or reload button [2].

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

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

Website: www.ijircce.com

Vol. 5, Issue 11, November 2017

Browser Level Collection: Second method of data collection is by modifying the browser. It shows the behaviour of single user over multiple sites. The data collection capabilities are enhanced by modifying the source code of existing browser. They provide much more versatile data as they consider the behaviour of single user on multiple sites [2].
Server Level Collection: Web server log [5] stores the behaviour of multiple users over single site. These log files can be stored in common log format or extended log format. Server logs are not able to store cached page views. Another technique used for usage data collection at server level is TCP/IP packet sniffing. Packet sniffers works by monitoring the net-work traffic and retrieve usage data directly.
Proxy Level Collection: Proxy servers are used by internet service provider to provide World Wide Web access to customers. These server stores the behaviour of multiple user at multiple site. These server functions like cache server and they are able to produce cached page views. By predicting the usage pattern of the visitor Web Usage Mining improves the quality of e- commerce services, personalizes the web [1] or enhances the performance of web structure and web server.
Server data are data that are collected from web servers; it includes log files, cookies and explicit user input. Servers contain different types of logs, which are considered to be the main date resource for web usage mining.
The most popular logs are:
 Common Log Format (CLF): created to keep track of requests that occur on a website in chronological order. It contains the IP address of the client, hostname, username, time stamp, file name and file size. CLF has the following elements:
 Remote host: the IP address or domain name of the client
 Base URL: the URL of the user request
 Date: the date and time of the request
 Method: the method used by the client, such as GET,HEAD or POST
 File: the file requested by the client
 Protocol: the protocol used
 Code: the status code of the three requests; it consists of 3 digits
 Bytes: the number of bytes returned to the client
 Referrer: the URL from the referring server
 User agent: the operating system type and version
The focus of related work to study and contrast the available technique to predict the web user behavior. Jagan and Rajagopalan [9] describe the web usage mining and algorithms used for providing personalization on the web. In this paper focused the data pre-processing and pattern analysis on the web and using the association rule mining algorithms.
Ladekar A. Pawar A. *et al*. [10] describe a web mining algorithm that aims at amending the interpretations of the draft's output of association rule mining. This algorithm is being tremendously used in web mining. The results obtained prove the robustness of the algorithm proposed in this paper.
Parvatikar S. and Joshi B. [11] this paper focused on Web Usage Mining is the user navigation patterns and their use of web resources. The different stages involved in this mining process and with the comparative analysis between the pattern discovery algorithms Apriori and FP-growth algorithm.
Deepa and Raajan [12] implemented the preprocessing techniques to convert the log file into user sessions which are suitable for mining and reduce the size of the session file by filtering the least requested pages using the preprocessing technique. Data Preprocessing is one of the important tasks before applying mining algorithms. It converts the raw log file into user session. In this work, we have briefly introduced log file preprocessing and implemented it in a CTI log file. Also, we produce the summary of the user session file. We have used filtering technique to remove least requested resources.
Anand N. [13] describes an internet usage details and provides them with the tools to understand the online behavior of their teenage children.
Singh A.P. and Jain R.C. [14] Different kinds of web usage mining techniques with their basic models and concepts are provided. In addition to that, for discovering the hidden patterns from web access log files a new model based on visual clustering is also suggested. The analysis of different methods of web usage mining.
Mishra R. and Choubey R. [15] describe the FPgrowth algorithm is obtaining a most frequently access paths and pages from the web log data and providing valuable information to user behavior.

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

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

Website: www.ijircce.com

Vol. 5, Issue 11, November 2017

Zubi Z. S. and Riani M.S.E. [16] discusses the use of web mining techniques is used to classify the web page's type according to user visits. This classification is helps to understand the web user behavior. The classification and association rule techniques for discovering the interesting
information from browsing patterns.

Avneet Saluja *et al.* [17] in their work is user future request prediction using web log records and user information. The purpose of the effort is to provide a benchmark for evaluating a various methods used in the past, a present and which can be used in a future to minimize the search time of a user on the network.

Web Usage Mining is an emerging field in research area. Many algorithms are used for Web Usage Mining in order to get better, accurate & efficient results such as Mehrdad Jalali *et al.* [8], Gang FANG *et al.* [9], Kobra Etminani *et al.* [10], Mamoun A. Awad *et al.* [11], Ashika Gupta *et al.* [12].

In [8], Mehrdad Jalali et al. gave the solution based on LCS algorithm for analyzing and process the user navigation patterns for next web page prediction. Their architecture has improved accuracy of classification & also it provides efficient online prediction . Some evaluation techniques also used for evaluating quality of the prediction found.

In [9], Gang FANG *et al.* proposed a double algorithm of Web usage mining based on sequence number, which is suitable for mining any session patterns in order to improve efficiency of presented algorithms and reduce the time of scanning database. They used the algorithm that turns session pattern of user into binary, and then uses up and down search strategy to double generate candidate frequent item sets. They also computed support by sequence number dimension in order to scan once session pattern of user, which is different from traditional double search mining algorithm. Their experiment indicates that efficiency of the algorithm is faster and more efficient than presented similar algorithms, such as, B_Apriori and B_ARDSM.

Kohonen's SOM (Self Organizing Map) model is applied to pre-processed web logs by Kobra Etminani *et al.* in [10] for clustering method. They used University's web server logs to extract the frequent patterns.

Markov Model is most widely known algorithm for Web Usage Mining. Mamoun A. Awad and Issa Khalil in [11] presented a new modified Markov model to overcome the issue of scalability in the number of paths. They also presented a new approach for creating classifier EC, which is based on two-tier prediction framework based on the training examples and the generated classifiers. Two-tier framework contributed to preserving accuracy (although one classifier was consulted) and reducing prediction time. The comparative results also show that large number of *N*-grams in the all-*K*th model does not always produce better prediction accuracy. Smaller *N*-gram models perform better than higher *N*-gram models in terms of accuracy.

One of the algorithms, which are very simple to use and easy to implement the Web Usage Mining task, is Apriori algorithm. Ashika Gupta *et al.* in their research work [12] emphasizes on web usage mining and has progress in web utilization with the help of web logs. The bonding of memory and time usage is compared by means of Apriori algorithm and improved Frequent Pattern Tree algorithm. But the main drawback of Apriori algorithm is that the candidate set creation is costly, if the data set is large and a long pattern is recognized. But FP-growth algorithm is not find good enough because it has lack of generating a good candidate method. Future research can combine FP-Tree with Apriori candidate generation method to solve the disadvantages of both apriori and FP-growth.

All these work for Web Usage Mining and recommendation is done to improve the accuracy and efficiency of the system. But still some performance issues are there. We present architecture and propose two prominent algorithms- k-means clustering algorithm and regression analysis. k-means is mostly used algorithm for clustering purpose and hence efficient too. Regression Analysis is an accurate method for prediction that applied on numeric values. Proposed architecture improves the accuracy and efficiency of prediction

Problem Definition

There are rich variants of browsing behaviour analysis techniques are available but most of them are suffers from the following issues:

1. Web server access log based technique only contains the partial user behaviour therefore need to improve the log management scheme
2. More than one pages are navigated in different times, therefore establishing the correlation between each user event and their corresponding web page is complex to learn by an algorithm
3. Huge data needs large time and space complexity
4. Inaccurate predictive methodology due to less number of feature availability on the user navigation pattern.

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

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

Website: www.ijircce.com

Vol. 5, Issue 11, November 2017

Limitations of Existing System:

1. Accuracy of system is quite less
2. Time consumption increase with increase in dataset size

Advantages of Proposed System:

1. Accuracy is high
2. Time consumption is very less as compared to previous systems
3. Classification accuracy is better than previous systems

Disadvantages of proposed system:

1. Doest not consider real time dataset
2. Processing speed depends on the machine configuration

Future Scope:

1. Can be implemented with other algorithms to check accuracy
2. Hybrid approach can also be implemented to improve accuracy
3. To be implemented using real world dataset

III. CONCLUSION

Web usage mining is indeed one of the emerging areas of research and important sub-domain of data mining and its techniques. In order to take full advantage of web usage mining and its all techniques, it is important to carry out preprocessing stage efficiently and effectively. This paper tries to deliver areas of preprocessing, including data cleansing, session identification, user identification. Once the preprocessing stage is well-performed, we can apply data mining techniques like clustering, association, classification etc for applications of web usage mining such as business intelligence, e-commerce, e-learning, personalization, etc. Web log mining is one of the recent areas of research in Data mining. Web Usage Mining becomes an important aspect in today's era because the quantity of data is continuously increasing. We deal with the web server logs which maintain the history of page requests.

REFERENCES

[1] J. Teevan, E. Adar, R. Jones, and M.A.S. Potts, "Information Re-Retrieval: Repeat Queries in Yahoo's Logs," Proc. 30th Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR '07), pp. 151-158, 2007.
[2] A. Broder, "A Taxonomy of Web Search," SIGIR Forum, vol. 36, no. 2, pp. 3-10, 2002.
[3] A. Spink, M. Park, B.J. Jansen, and J. Pedersen, "Multitasking during Web Search Sessions," Information Processing and Management, vol. 42, no. 1, pp. 264-275, 2006.
[4] R. Jones and K.L. Klinkner, "Beyond the Session Timeout:
Automatic Hierarchical Segmentation of Search Topics in Query Logs," Proc. 17th ACM Conf. Information and Knowledge Management (CIKM), 2008. P. Boldi, F. Bonchi, C. Castillo, D. Donato, A. Gionis, and S. Vigna, "The Query-Flow Graph: Model and Applications," Proc. 17th ACM Conf. Information and Knowledge Management (CIKM), 2008.
[6] D. Beeferman and A. Berger, "Agglomerative Clustering of a Search Engine Query Log," Proc. Sixth ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (KDD), 2000.
[7] R. Baeza-Yates and A. Tiberi, "Extracting Semantic Relations from Query Logs," Proc. 13th ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (KDD), 2007.
[8] J. Han and M. Kamber, Data Mining: Concepts and Techniques. Morgan Kaufmann, 2000.
[9] W. Barbakh and C. Fyfe, "Online Clustering Algorithms," Int'l J. Neural Systems, vol. 18, no. 3, pp. 185-194, 2008.
[10] Lecture Notes in Data Mining, M. Berry, and M. Browne, eds. World Scientific Publishing Company, 2006.
[11] V.I. Levenshtein, "Binary Codes Capable of Correcting Deletions, Insertions and Reversals," Soviet Physics Doklady, vol. 10, pp. 707-710, 1966.
[12] M. Sahami and T.D. Heilman, "A Web-based Kernel Function for Measuring the Similarity of Short Text Snippets," Proc. the 15th Int'l Conf. World Wide Web (WWW '06), pp. 377-386, 2006.
[13] J.-R. Wen, J.-Y. Nie, and H.-J. Zhang, "Query Clustering Using User Logs," ACM Trans. in Information Systems, vol. 20, no. 1, pp. 59-81, 2002.
[14] A. Fuxman, P. Tsaparas, K. Achan, and R. Agrawal, "Using the Wisdom of the Crowds for Keyword Generation," Proc. the 17th Int'l Conf. World Wide Web (WWW '08), 2008.
[15] K. Avrachenkov, N. Litvak, D. Nemirovsky, and N. Osipova, "Monte Carlo Methods in PageRank Computation: When One Iteration Is Sufficient," SIAM J. Numerical Analysis, vol. 45, no. 2, pp. 890-904, 2007.