ISSN 2395-1621

# Survey On Web Mining For User's Behavior Information from Central Database System

[#1]Miss. Poonam S. Jadhavar, [#2]Prof. Varsha Dange

[1]jadhavar.poonam9@gmail.com
[2]dange.varshar@gmail.com

[#1]M.E. Student, Dept. of Computer Engineering, DPCOE, Pune
Pune, India

[#2]Assistant Professor, Dept. of Computer Engineering, DPCOE, Pune
Pune, India

**ABSTRACT**

**The Web mining extracts useful information from the web pages. Web mining techniques seek to extract knowledge from Web data, including web documents, hyperlinks between documents, and usage logs of web sites. Web usage mining mines knowledge from diverse websites. Extracting appropriate data from deep web pages is an exigent dilemma due to the overflow of data in to the web. Web servers generate a huge amount of information on web users browsing activities. These are called click stream or web access log data. The click stream data can be enriched with information about the content of visited pages. Web usage mining extract knowledge from the web log file but the problem is these log files are private and the companies does not share their private information to public. An Application Program Interface (API), we have developed to share these private log files.**

ARTICLE INFO

**Article History**

Received: 4th December 2016

Received in revised form :

4th December 2016

Accepted: 8th December 2016

**Published online :**

**8th December 2016**

**Keywords: Tool for Usage Mining, Web Usage Mining, Web Log Mining, Free Log.**

## I. INTRODUCTION

A lot of research has been done on Web Usage Mining. When the user browses the web pages, user leaves some valuable information in web log. Web usage mining automatic discovers the knowledge from the data collected in log file. These log files are extremely important. This log information is useful for target potential customers for electronic commerce, Enhancement of the quality and delivery of internet information services to the end user, Improvement of web server system performance, Identification of potential prime advertisement locations, Facilitates personalization of sites, Improvement of site design, Fraud/ Intrusion detection and prediction of user' s actions (allows pre-fetching). The numbers of companies gather users' behavior information in the form of log files but the problem is they don't share this information with rest of the world. We have developed and implemented a system which store users' behavior information in central database. This system could be helpful to all the people to store their behavior information and retrieve their and others' behavior information. The new concept of this work is to offer everyone the same log information that important web companies have. To implement this system we have developed a tool, which transfers the users' behavior log information to the Local log database. From there it is transferred to the central database and then it is available for public access.

## II.WEB MINING TAXONOMY

Web Mining can be broadly divided into three distinct categories according to the kinds of data to be mined. Following figure 1 explains it best.

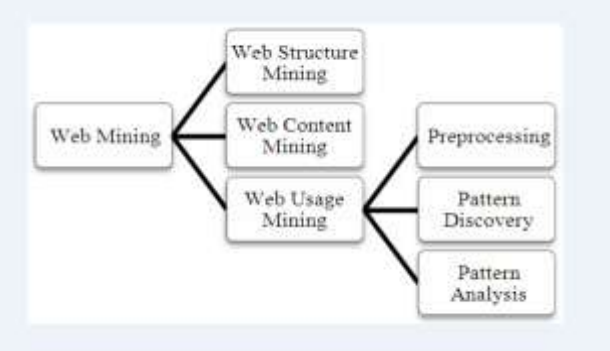


Fig 1. Web Mining Taxonomy

 

**A. Web Content Mining**

Web Content Mining is the process of extracting useful information from the contents of Web documents. Content data corresponds to collection of facts a Web page was designed to convey to the users. It may consist of text, images, audio, video, or structured records such as lists and tables. Text mining and its application to Web content has been the most widely researched. Research activities in this field also involve using techniques from AI such as Information Retrieval [IR], Natural Language Processing [NLP], and Image processing and computer vision.

**B. Web Structure Mining**

The structure of a typical Web graph consists of Web pages as nodes, and hyperlinks as edges connecting between two related pages. Web Structure Mining can be regarded as the process of discovering structure information from the Web. This type of mining can be further divided into two kinds based on the kind of structural data used.

• Hyperlinks: A Hyperlink is a structural unit that connects a Web page to different location, either within the same Web page or to 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: 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.

**C. Web Usage Mining**

Web Usage Mining is the application of data mining techniques to discover interesting usage patterns from Web data, in order to understand and better serve the needs of Web-based applications. Usage data captures the identity or origin of Web users along with their browsing behavior at a Web site. Capturing, Modeling and analyzing of behavioral patterns of users is the goal of this web mining category. Web usage mining process can be divided into three independent tasks: Pre-processing, Pattern discovery and pattern analysis.

The following Figure 2 shows this process

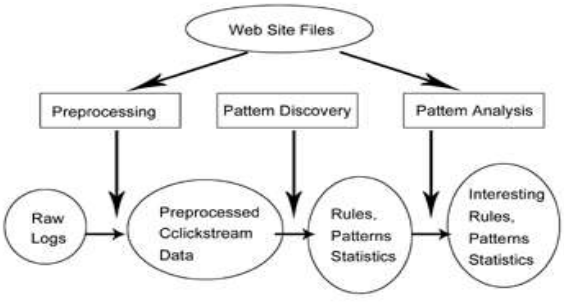


Fig 2. Web mining process

## III. MOTIVATION AND RELATED WORK

Millions of users access web sites in all over the world. When they access a websites, a large amount of data generated in log files which is very important because many times user repeatedly access the same type of web pages and the record is maintained in log files. These series can be considered as a web access pattern which is helpful to find out the user behavior. Through this behavior information, we can find out the accurate user next request prediction that can reduce the browsing time of web pages. In recent years, there has been an increasing number of research works done with regard to web usage mining 'Future request prediction'. The main motivation of this survey is to know the research has been done on Web usage mining in future request prediction. [1] Proposed algorithm automatically discovers pages in a website whose location is different from where visitors expect to find them. For that purpose they use 'Backtrack' as a key and this key is used for algorithm from the point where the users backtrack. This is the expected location for the page. One more algorithm that selects the sets of navigation links to optimize visitor time or benefit of the web site. [2] investigate a 'real time management engine' with the help of historical data and online visitation patterns of e-commerce site visitors. [3] described and compared three different web usage mining techniques, based on transaction clustering, usage clustering and association clustering. Usage Clustering and Association rule discover to extract a usage knowledge for web personalization to recommendation of pages.

## IV. LITERATURE SURVEY

Millions of users access web sites in all over the world. When they access a websites, a large amount of data generated in log files which is very important because many times user repeatedly access the same type of web pages and the record is maintained in log files. These series can be considered as a web access pattern which is helpful to find out the user behavior. Through this behavior information, we can find out the accurate user next request prediction that can reduce the browsing time of web pages. In recent years, there has been an increasing number of research works done with regard to web usage mining 'Future request prediction'. The main motivation of this survey is to know the what research has been done on Web usage mining in future request prediction.

**Ramakrishnan Srikant, Yinghui Yang "Mining Web Logs to Improve Website Organization", Proceedings of the 10th international conference on World Wide Web, 2001**

Proposed algorithm automatically discovers pages in a website whose location is different from where visitors expect to find them. For that purpose they use 'Backtrack' as a key and this key is used for algorithm from the point where the user backtrack. This is the expected location for the page. One more algorithm that selects the sets of navigation links to optimize visitor time or benefit of the web site.

**Debra Vander Meer, Kaushik Dutta, Anindya Datta "Enabling Scalable Online Personalization on the Web" Published in: Proceeding of the 2nd ACM conference on Electronic Commerce, October 17-20, 2000**

 

Investigate a 'real time management engine' with the help of historical data and online visitation patterns of e-commerce site visitors.

**Bamshad Mobasher, Robert Cooley, Jaideep Srivastava "Automatic Personalization Based on Web Usage Mining" Published in: Magazine Communication of the ACM, Volume 43 Issues 8, Aug 2000**

Described and compared three different web usage mining techniques, based on transaction clustering, usage clustering and association clustering. Usage Clustering and Association rule discover to extract a usage knowledge for web personalization to recommendation of pages.

**Mathias Gery, Hatem Haddad "Evaluation of Web Usage Mining Approaches for User‟s Next Request Prediction" WIDM‟03 Proceedings of the 5th ACM international workshop on web information and data management p.74-81, November 7-8,2003.**

According to Author distinguished three web mining approaches that exploit web logs: Association Rules (AR), Frequent Sequences (FS) and Frequent Generalized Sequences (FGS). Algorithms for three approaches were developed and experiments have been done with real web log data.

Association Rule: In data mining, association rule learning is a popular and well researched method for discovering interesting relations between variables in large database. Describes analyze and present strong rules discovered in database using different measures of interestingness. The problem of finding web pages visited together is similar to finding associations among item sets in transaction databases. Once transaction have been identified each of them could represent a basket and each research an item.

Frequent Sequences: The attempt of this technique is to discover time ordered sequences of URLs that have been followed by past users.

Frequent Generalized Sequences (FGS): a generalized sequence is a sequence allowing wildcards in order to reflect the users navigation in a flexible way. In order to extract frequent generalized subsequences they have used the generalized algorithm proposed by Gaul.

Author performed some experiments for this purpose they used three collections of web log datasets. One weblog dataset for small web site, another for large website and the third weblog dataset for intranet website. By using above three web mining approaches they evaluate the three different types of real web log data and they found Frequent Sequence (FS) gives better accuracy than AR and FGS.

**Enrique Frıas-Martınez, Vijay Karamcheti "A Customizable Behaviour Model for Temporal A Customizable Behaviour Model for Temporal" WEBKDD 2002 Mining web data for discovering usage patterns and profiles, p.66-85, 2003**

Present a model that constructed sequential rules which are unlike clustering and association rule, here they capture the sequentially and temporality in which web pages are visited. To maintain sequentiality the rules maintain the sequence of the click stream of the antecedent and of the consequent. The concept of temporality is reflected with distance between antecedent and the consequent measured by the number of user clicks to go from one page to another. The concept of antecedent and the consequent is the important for the prediction system because it allows the rules to express not only what pages are going to be accessed but also precisely when they are going to be accessed. They proposed customizable prediction system, this means that the prediction system can be adapted, depending on the characteristics of the server (number of pages, architecture of the server, number of links per page etc.) in order to more accurately capture the behavior of its users.

**A. Anitha, "A New Web Usage Mining Approach for Next Page Access Prediction", International Journal of Computer Applications, Volume 8– No.11, October 2010**

Proposed to integrate Markov model based sequential pattern mining with clustering. With the help of proposed approach approximately 12% of prediction accuracy increases compared to traditional Markov model. The main advantage of proposed hierarchical clustering approach is that every object must be candidate of only one cluster.

| Author | Method | Application | Publication Year |
|---|---|---|---|
| Alexandra's, Dimitris, Yannis [2] | Web page perfecting in to cache | Prediction enabled web server | 2001 |
| Mathias Gery, Hatem Haddad [6] | Associate rule (AR), Frequent sequences (FS) and Frequent generalized sequence(FGS) | Evaluate AR, FS, FGS | 2003 |
| Enrique, Vijay Karamcheti [5] | Clustering and Sequential associate rule | Behavior Model | 2003 |
| Nien-Yijan, Nancy P. Lin[9] | Trend based prediction | Prediction System Architecture | 2007 |
| Mehrdad, Norwati Ali, Md Nasir[7] | Online/Offline phase of architecture, LCS Algorithm, clustering | Online prediction future user moments | 2008 |
| Mehrdad, Norwati Ali, Md Nasir[8] | Online/Offline phase of architecture, LCS Algorithm, clustering | WebPUM | 2010 |
| A. Anitha[1] | Sequential pattern mining with kth order Markov model clustering | Next Page Access Prediction system | 2010 |

## V. PROPOSED SYSTEM

The objective of a project is to build a web mining tool, which will act as an Application program interface to share user behavior information for different web sites. Conceptual framework for tool is shown in figure 3 .

**Central Web Mining Public Server**

Public Server Is designed to make available log information to general public. All the people can download log data from the central server. Central database maintain the all website log data which is uploaded by the different user.

**Local Server**

It can be one computer or local server of internet cafe, college, university, Industry, Organization, Corporate organization. From the local server, user can upload his own behavior information or others behavior information on the public server. People can upload this information to serve the society by providing their log information to community. All the people get equal opportunity to be at the same commercial value by having personal log information that important web companies have.

 

**Local Log Manager**

We have designed a local log manager module to view collected log information. With the help of this module, User can filter the log files. If he does not want to share these log files, he can store these log files to his personal database. This information can be useful to him for analysis or research purpose. If the user wants to upload behavior information by his name, he can upload by name or by being anonymous user on Public server.

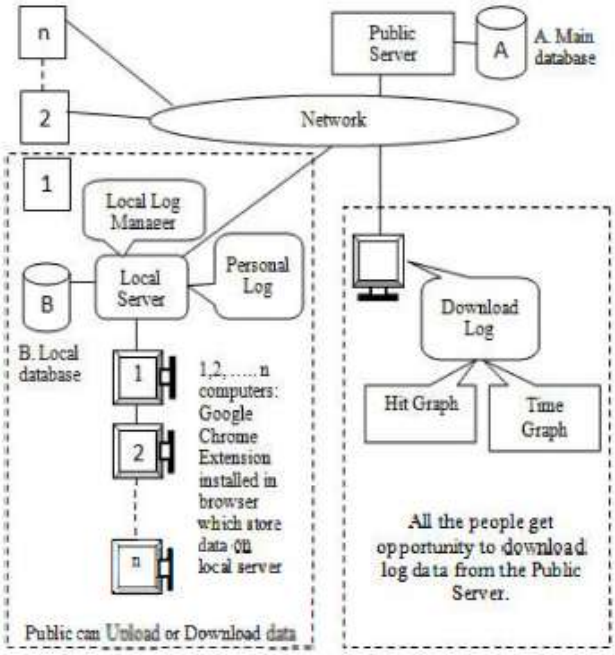


FIG. 3: CONCEPTUAL FRAMEWORKS FOR TOOL

**Personal Log Module**

With this module, User could view the stored log files in personal database. With this module user can edit or delete the log record from the same personal database.

**1, 2 ... n Computers**

Different users usually surf web pages on browser. We have designed a tool which is nothing but the Extension. This tool sends the log information on the local database when users browse the web pages. From local database it can be uploaded to central web mining public server.

**Downloading Log**

Anyone can download log information from the public server. In two ways it is possible to download the log from the public server. One way is using a Download Log module. We have designed a Download log module with the help of this module user can download the logs information. The module also provides functionality, that the user could see the analysis in graphical form. Like Hits Graph, Time Graph (different timing required to fetch the web pages). Another way is Java Server Web pages, with the help of these web pages; user can view if they wanted to see response times, response sizes, number of requests for different pages. People can categorize the users according to Location, country, region, and city and also represent this information graphically on Google Chart.

## VI. ALGORITHM

Hyperlink-Induced Topic Search (HITS; also known as hubs and authorities) is a link analysis algorithm that rates Web pages, developed by Jon Kleinberg. Kleinberg gives two forms of web pages called as hubs and authorities. Hubs are the pages that act as resource lists. Authorities are pages having important contents. A good hub page is a page which is pointing to many authoritative pages on that content and a good authority page is a page which is pointed by many good hub pages on the same content. A page may be a good hub and a good authority at the same time The HITS algorithm treats WWW as directed graph G(V,E), where V is a set of vertices representing pages and E is set of edges corresponds to link. Figure 1 shows the hubs and authorities in web.

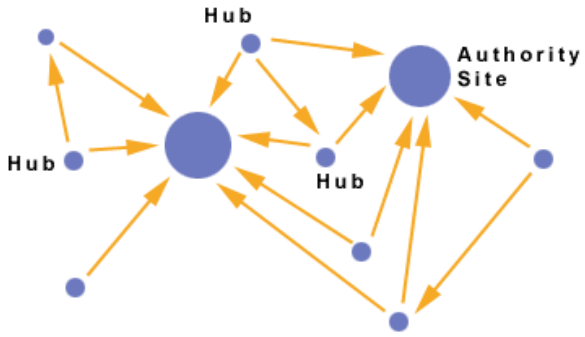


Fig.4: Hubs And Authorities

It has two steps:

1. Sampling Step:- In this step a set of relevant pages for the given query are collected.
2. Iterative Step:- In this step Hubs and Authorities are found using the output of sampling step.

Following expressions are used to calculate the weight of Hub (Hp) and the weight of Authority (Ap).

$$H_p = \sum_{q \in I(p)} A_q$$

$$A_p = \sum_{q \in B(p)} H_q$$

here Hq is Hub Score of a page, Aq is authority score of a page, I(p) is set of reference pages of page p and B(p) is set of referrer pages of page p, the authority weight of a page is proportional to the sum of hub weights of pages that link to it. Similarly a hub of a page is proportional to the sum of authority weights of pages that it links to.

**Constraints with HITS algorithm Following are some constraints of HITS algorithm.**

- Hubs and authorities: It is not easy to distinguish between hubs and authorities because many sites are hubs as well as authorities.
- Topic drift: Sometime HITS may not produce the most relevant documents to the user queries because of equivalent weights

 

- Automatically generated links: HITS gives equal importance for automatically generated links which may not have relevant topics for the user query
- Efficiency: HITS algorithm is not efficient in real time.

HITS was used in a prototype search engine called Clever for an IBM research project. Because of the above constraints HITS could not be implemented in a real time search engine.

## VII. CONCLUSION

In this paper we proposed a new Web Mining Service which can make avail private log file of different website to general public. We implemented tool which collect private log file when user browsing and stores on central web mining server . This log information will be helpful for their business. Which improve their website and to get equal commercial value like other top rated web sites Using HITS algorithm. Through this mining project we want to show how these mining API can be useful to all people.

## REFERENCES

[1] A. Anitha, "A New Web Usage Mining Approach for Next Page Access Prediction", International Journal of Computer Applications, Volume 8– No.11, October 2010

[2] Alexandros Nanopoulos, Dimitris Katsaros and Yannis Manolopoulos "Effective prediction of web-user accesses: A data mining approach," in Proc. Of the Workshop WEBKDD, 2001.

[3] Bamshad Mobasher, Robert Cooley, Jaideep Srivastava "Automatic Personalization Based on Web Usage Mining" Published in: Magazine Communication of the ACM, Volume 43 Issues 8, Aug 2000

[4] Debra Vander Meer, Kaushik Dutta, Anindya Datta "Enabling Scalable Online Personalization on the Web" Published in: Proceeding of the 2nd ACM conference on Electronic Commerce, October 17-20, 2000

[5] Enrique Fr´ıas-Mart´ınez, Vijay Karamcheti "A Customizable Behaviour Model for Temporal A Customizable Behaviour Model for Temporal" WEBKDD 2002 Mining web data for discovering usage patterns and profiles, p.66-85, 2003

[6] Mathias Gery, Hatem Haddad "Evaluation of Web Usage Mining Approaches for User‟s Next Request Prediction" WIDM‟03 Proceedings of the 5th ACM international workshop on web information and data management p.74-81, November 7-8,2003.

[7] Mehrdad Jalali,Norwati Mustapha, Ali Mamat , Md. Nasir B Sulaiman "A.

 