Analysis of Web Log Mining

ISSN 2395-1621

[#1]Rutuja A Suryawanshi, [#2]Supriya H Wakchaure, [#3]Sonal S Patil,
[#4]Ashish S More, [#5]Prof. Rashmi Tundalwar

[4]ashishmore91@gmail.com

[#12345]Department of Computer Engineering,

Dhole Patil College of Engineering, Pune,
Savitiribai Phule Pune University India.

ABSTRACT

In this the Existing model study the method to extract the user sessions from the given log files. Initially, each user is identified according to his/her IP address specified in the log file and corresponding user sessions are extracted. Two types of logs i.e., server-side logs and client-side logs are commonly used for web usage and usability analysis. Server-side logs can be automatically generated by webservers, with each entry corresponding to a user request. Client-side logs can capture accurate, comprehensive usage data for usability analysis. In this existing paper process includes 3 stages, namely Data cleaning, User identification, and Session identification. In this paper implementing these three phases. Depending upon the frequency of users visiting each page mining is performed. By finding the session of the user we can analyse the user behaviour by the time spend on a particular page.

ARTICLE INFO

Article History

Received: 10th October 2016

Received in revised form :

10th October 2016

Accepted: 18th October 2016

Published online :

18th October 2016

Keywords: Data mining, Weblog Mining, Session log.

I. INTRODUCTION

As Statistical Analysis and Web Usage Mining are two ways to analyze users' web browsing behavior. The result of Statistical Analysis contains Page Views, Page Browsing Time, and so on. Web usage mining applies data mining methods to discover web usage pattern through web usage data. Item-Set Mining, Sequential Pattern Mining, and Graph Mining, are examples of data mining methods that can be used to analyze web usage data [7-8].

Web usage data can be collected from three sources: server level, client level, and proxy level [9]. Statistical analysis and web usage mining usually use Server Log as main data source which is server level data source. Server log is a file that automatically created by web server and kept on server. It contains some data about requests which are sent to web server. During reconstruction of user's session, server log may not be fully reliable because in some cases such as page cashing and POST method, data are not recorded in server log [9].

1.1 Web Content Mining:

Web Content Mining is the process of extracting useful information from the contents of Web documents. Content data corresponds to the 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. Some of the research issues addressed in text mining are, topic discovery, extracting association patterns, clustering of web documents and classification of Web Pages.

1.2 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: In addition, the content within a Web page can also be organized in a tree-

© 2016, IERJ All Rights Reserved

structured format, based on the various HTML and XML tags within the page.

1.3 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. Web usage mining itself can be classified further depending on the kind of usage data considered:

- **Web Server Data:** They correspond to the user logs that are collected at Web server. Some of the typical data collected at a Web server include IP addresses, page references, and access time of the users.

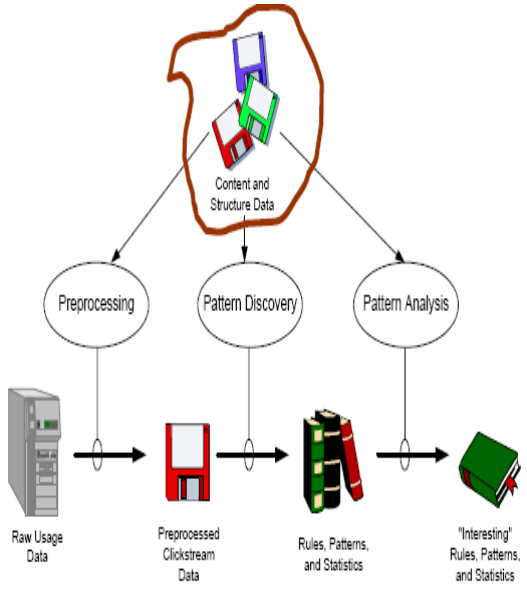

Fig 1. Web Mining Process

II. LITERATURE SURVEY

[1] In this paper authors developed a new method for the identification and improvement of navigation-related Web usability problems by checking extracted usage patterns against cognitive user models. As demonstrated by our case study, our method can identify areas with usability issues to help improve the usability of Web systems. Once a website is operational, our method can be continuously applied and drive ongoing refinements. In contrast with traditional software products and systems, Web based applications have shortened development cycles and prolonged maintenance cycles. In this method can contribute significantly to continuous usability improvement over these prolonged maintenance cycles. The usability improvement in successive iterations can be quantified by the progressively better effectiveness (higher task completion rate) and efficiency (less time for given tasks).

[2] In this paper author implemented Web usage mining process of extracting useful information from web server logs based on the browsing and access patterns of the users. The information is especially valuable for business sites in order to achieve improved customer satisfaction. Based on the user's needs, Web Usage Mining discovers interesting usage patterns from web data in order to understand and better serve the needs of the web based application. Web Usage Mining is used to discover hidden patterns from weblogs. It consists of three phases like Pre-processing, pattern discovery and Pattern analysis. In this paper, we present each phase in detail, the process of extracting useful information from server log files and some of application areas of Web Usage Mining such as Education, Health, Human-computer interaction, and Social media.

[3] In this paper study on Web mining is a technology that has strong practical significance in E-commerce. This technology not only helps enhance the performance of the website and understand customer need, but also serves as a basis for enhancing the topology of the site and hyperlinks. This research discusses the classification of web mining, which is classified into three categories: web content mining, web structural mining, and web usage mining. The web mining process in terms of how to mine a large number of data to obtain customer behavior information is also discussed. It also discusses the importance of the application of web mining to E-commerce, which has a major influence to the merchant, customer, and company.

[4] This paper proposes a novel method to efficiently provide better Web-page recommendation through semantic-enhancement by integrating the domain and Web usage knowledge of a website. Two new models are proposed to represent the domain knowledge. The first model uses an ontology to represent the domain knowledge. The second model uses one automatically generated semantic network to represent domain terms, Web-pages, and the relations between them. Another new model, the conceptual prediction model, is proposed to automatically generate a semantic network of the semantic Web usage knowledge, which is the integration of domain knowledge and Web usage knowledge.

[5] In this paper, he conduct extensive analyses and comparisons to evaluate the effectiveness of task trails in several search applications: determining user satisfaction, predicting user search interests, and suggesting related queries. Experiments on large scale data sets of a commercial search engine show that: (1) Task trail performs better than session and query trails in determining user satisfaction; (2) Task trail increases webpage utilities of end users comparing to session and query trails; (3) Task trails are comparable to query trails but more sensitive than session trails in measuring different ranking functions; (4) Query terms from the same task are more topically consistent to each other than query terms from different tasks; (5) Query suggestion

© 2016, IERJ All Rights Reserved

based on task trail is a good complement of query suggestions based on session trail and click-through bipartite. The findings in this paper verify the need of extracting task trails from web search logs and enhance applications in search and recommendation systems.

[6] This paper proposes a fully automated information extraction methodology for weblogs. The methodology integrates a set of relevant approaches based on the use of web feeds and processing of HTML for the extraction of weblog properties. The approach includes a model for generating a wrapper that exploits web feeds for deriving a set of extraction rules automatically. Instead of performing a pairwise comparison between posts, the model matches the values of the web feeds against their corresponding HTML elements retrieved from multiple weblog posts.

III. EXISTING MODEL

In this the Existing model study the method to extract the user sessions from the given log files. Initially, each user is identified according to his/her IP address specified in the log file and corresponding user sessions are extracted. Two types of logs ie., server-side logs and client-side logs are commonly used for web usage and usability analysis. Server-side logs can be automatically generated by webservers, with each entry corresponding to a user request. Client-side logs can capture accurate, comprehensive usage data for usability analysis. In this existing paper process includes 3 stages, namely Data cleaning, User identification, Session identification. In this paper implementing these three phases. Depending upon the frequency of users visiting each page mining is performed. By finding the session of the user we can analyse the user behaviour by the time spend on a particular page.

IV. PROPOSED WORK

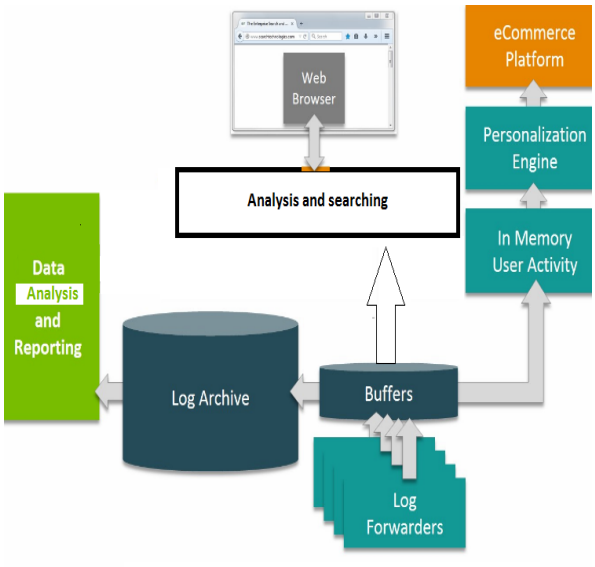

Fig 2. System architecture

- Product View
  - Occurs every time a product is displayed on a page view
  - Typical Types: Image, Link, Text
- Product Click-through
  - Occurs every time a user “clicks” on a product to get more information
- Maximum purchases of particular product
- And available offers.
- Mgs and mail through offers are informed.

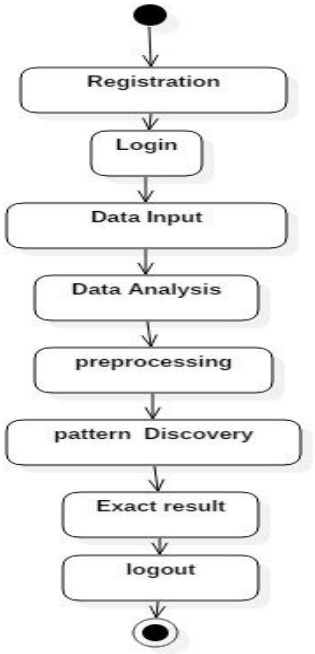

Fig 3 Flow of system

Proposed Algorithm:

Web Log Mining algorithm

Input : Page P.
Output : Storage of user activity on session log.
Begin
// Segment the page
Ψ = buildSegments(P)
Repeat until page is closed
begin
locate the mouse pointer (x,y)
seg_id = the segment for the (x,y)
enTime = entry time into the segment;
exTime = exit time from the segment;
update session log (seg_id, enTime,exTime);
End
End

Mail generation algorithm

Input : Mail data.
Output :Send mail.
Begin
// Mail server access
smtpHostServer ();
Mail id input();

© 2016, IERJ All Rights Reserved

```
Repeat until the mail send
begin
Get properties();
Getinstance();
//Access email utility
 EmailUtil();
End
End
```

V. CONCLUSION

Web usage mining has emerged as the essential tool for realizing more personalized user-friendly and business-optimal Web services. Traditionally, Web usage mining is used by e-commerce sites and other organization's to organize their sites and to increase profits.

REFERENCES

[1] Ruili Geng, and Jeff Tian "Improving Web Navigation Usability by Comparing Actual and Anticipated Usage"IEEE TRANSACTIONS ON HUMAN-MACHINE SYSTEMS, VOL. 45, NO. 1, FEBRUARY 2015.

[2] G. Neelima and Sireesha Rodda, "An Overview on Web Usage Mining", Springer International Publishing Switzerland December 2015.

[3] Gan Teck Wei, Shirly Kho, Wahidah Husain, Zurinahni Zainol " A Study of Customer Behaviour Through Web Mining"Volume 2, Issue 1 available at www.scitecresearch.com/journals/index.php/jisct/index, February, 2015.

[4] Thi Thanh Sang Nguyen, Hai Yan Lu, and Jie Lu, "Web-Page Recommendation Based on Web Usage and Domain Knowledge", IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 26, NO. 10, OCTOBER 2014.

[5] Zhen Liao, Yang Song, Yalou Huang, Li-weiHe, and Qi He "Task Trail: An Effective Segmentation of User Search Behavior", IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 26, NO. 12, DECEMBER 2014.

[6] George Gkotsis • Karen Stepanyan • Alexandra I. Christie • Mike Joy,"Entropy-based automated wrapper generation for weblog data extraction", Received: 31 October 2012 / Revised: 24 October 2013 Accepted: 4 November 2013 / Published online: 21 November 2013, Springer Science+Business Media New York 2013.

[7.] Buchner, A. and Mulvenna, M. D., Discovering internet marketing intelligence through online analytical Web usage mining. SIGMOD Record, (4) 27, 1999.

[8.] Chen, M. S., Park, J. S., and Yu, P. S., Data mining for path traversal patterns in a Web environment. In Proceedings of 16th International Conference on Distributed Computing Systems, 1996.

[9] Agrawal, R. and Srikant, R., Fast algorithms for mining association rules. In Proceedings of the 20th VLDB conference, pp. 487-499, Santiago, Chile, 1994.

[10] Han, E-H, Boley, D., Gini, M., Gross, R., Hastings, K., Karypis, G., Kumar, V., and Mobasher, B., More, J., Document categorization and query generation on the World Wide Web using WebACE. Journal of Artificial Intelligence Review, January 1999.

[11] Herlocker, J., Konstan, J., Borchers, A., Riedl, J.. An algorithmic framework for performing collaborative filtering. To appear in Proceedings of the 1999 Conference on Research and Development in Information Retrieval, August 1999.

[12] Han, E-H, Karypis, G., Kumar, V., and Mobasher, B., Clustering based on association rule hypergraphs. In Proccedings of SIGMOD'97 Workshop on Research Issues in Data Mining and Knowledge Discovery (DMKD'97), May 1997.

[13]Han, E-H, Karypis, G., Kumar, V., and Mobasher, B., Hypergraph based clustering in highdimensional data sets: a summary of results. IEEE Bulletin of the Technical Committee on Data Engineering, (21) 1, March 1998.

[14]Konstan, J., Miller, B., Maltz, D., Herlocker, J., Gordon, L., and Riedl, J., GroupLens: applying collaborative filtering to usenet news. Communications of the ACM (40) 3, 1997.

[15]Spiliopoulou, M. and Faulstich, L. C., WUM: A Web Utilization Miner. In Proceedings of EDBT Workshop WebDB98, Valencia, Spain, LNCS 1590, Springer Verlag, 1999.

[16]Schechter, S., Krishnan, M., and Smith, M. D., Using path profiles to predict HTTP requests. In Proceedings of 7th International World Wide Web Conference, Brisbane, Australia, 1998.

[17]Nasraoui, O., Frigui, H., Joshi, A., Krishnapuram, R., Mining Web access logs using relational competitive fuzzy clustering. To appear in the Proceedings of the Eight International Fuzzy Systems Association World Congress, August 1999.

[18.]Perkowitz, M. and Etzioni, O., Adaptive Web sites: automatically synthesizing Web pages. In Proceedings of Fifteenth National Conference on Artificial Intelligence, Madison, WI, 1998.

[19]Shardanand, U., Maes, P., Social information filtering: algorithms for automating "word of mouth."In Proceedings of the ACM CHI Conference, 1995.

© 2016, IERJ All Rights Reserved