Volume 5, Issue 5, May 2015 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

Hidden Data and Key Term Extraction from Known ECLF for Identifying User Behavior

P. Kavitha[1], Dr. G. N. K.Suresh Babu[2]

[1]Research Scholar, Department of Computer Science, Bharathiar University, Coimbatore, Tamil Nadu, India

[2]Associate Professor, GKM College of Engineering and Technology, Chennai, Tamil Nadu, India

Abstract: In today's world, World Wide Web is the space that is used by company, academic and non academic people for gathering information they need. Common people rely simply on search engines for the information they need. Many search engines trying to provide result which is required for the user using recommending system. Since nothing could meet users need at most of the cases. For a single query lot of information are provided by search engine and in turn search engine provide result which has more hit rate. The result is not so relevant even though search engines are using personalization technique for learning specific need by users, since in most cases users avoid being personalized for security reasons. The proposed system uses Extended Common Log Files for extracting hidden information about a user for search engines to provide better result for each and every user.

Keywords: Extended Common Log Files, Hit rate, Personalization, Recommending system, Search engine.

I. INTRODUCTION

Web mining is the area which uses Data mining system. Web sites have a lot of consistent web pages that are developed and maintained by an association. This is the most direct way that an organization can learn about their current and potential customers. The companies can study visitor's activities through web analysis, and find the patterns in the visitor's behavior. Web mining is the process of discovering and analyzing the useful information from the World Wide Web.

The Web can be viewed as the largest unstructured data source available, although the data on the Web sites, which composed them, is structured. This presents an exigent task for effective design and access to Web pages. Web mining is a term used for applying data mining techniques to Web access logs. It is useful in personalized web pages, for web search, user tracking, understanding of user behavior, decision making etc. An important constituent category of Web Mining is Web Log mining also known as Web Usage mining, is the process of extracting interesting patterns from web access logs [1].

There are several general challenges associated with obtaining due results from the data. It is mainly because of irrelevant information is mixed with useful one. Secondly, multiple server requests may be generated by a single user action. Thirdly, multiple user actions may generate the same server request [2] and it is difficult for the system to learn more about user due to security issues like firewall, proxies etc.

The paper is organized as follows: Section 2 involves in analyzing web logs related works. Section 3 involves in extracting useful data's from user access content using TF-IDF algorithm and method for obtaining needed contents from web logs for identifying specific user need. Section 4 involves in extracting hidden data about a user from known Extended Common Log Files (ECLF). Section 5 focus on the future prospects and conclusion.

II. RELATED WORKS

Weblogs are the plaintext files contain information about user IP Address, User Name, URL that Referred, Access Request, Time stamp, error codes etc. and those are generally reside in the web servers. Server logs are delegated Transfer log, Error Log, Agent log, and Referrer Log. This section illustrates the methodologies which have been used in previous research work on user behavior analysis in web usages.

Chen, Chau and Tseng (2009) studied that it is vital for students to acquire knowledge and hands-on experience in Web mining. In this paper, design a Web mining application using the open Web APIs provided by Google, Amazon, and eBay[3]. Claudia Elena (2011) [4] used data mining techniques and researches that click stream or web access log data better understand and characterize web users. The goal of their project is to analyse user behaviour by mining enriched web access log data.

A conceptual schema can be created [6] that can describe the semantics of a large volume of unstructured web data to manage them [5]. Jian Pei, Jiawei Han, B. Mortazavi-asl & Hua Zhu studied the problem of mining access pattern from web logs efficiently. A data structure called web access pattern tree or WAP tree in short is developed for efficient mining of access pattern from the web logs [7].

Sudharsan.N.S et al [8] proposed a method to detect publishing site by considering key terms identified using TF-IDF algorithm from given webpage. TF-IDF, term frequency–inverse document frequency, is a numerical statistic which reflects how important a word is to a document in a collection or corpus. It is often used as a weighting factor in information and text mining.

Sudharsan.N.S et al [9] proposed a method to identify poisonous site in online with help of mining its Uniform Resource Locator (URL) pattern. Here the structure of URL is mined from the given page for identifying abnormal feature. DeMin Dong describes about various exploring technique on Web Usage Mining and its Application [10] and Renata Ivancsy et al describes about various method of analyzing web log patterns for web user identification [11].

III. PROPOSED WORK

The proposed system involves in identifying factors which is used to identify specific user. Fig.1. explains the overall architecture of the proposed system. Web logs are nothing but files which will be stored in the web server which contain user activities such as what website they access, what link they have clicked, time at which when the user clicked the link etc.

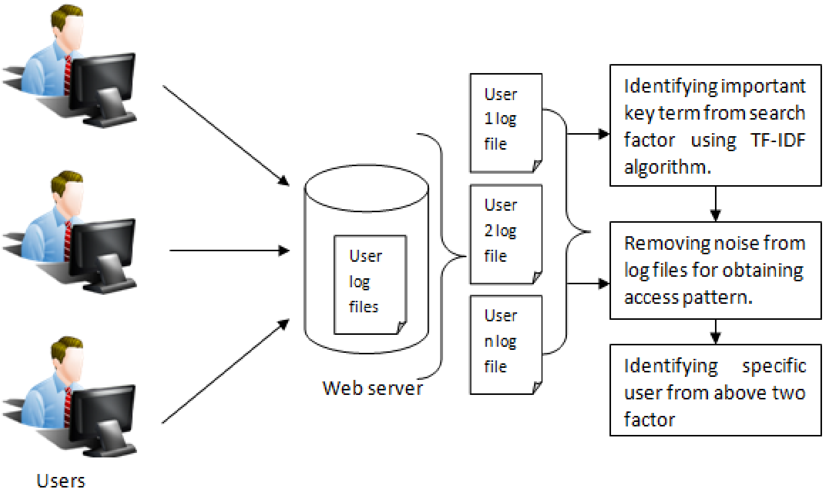
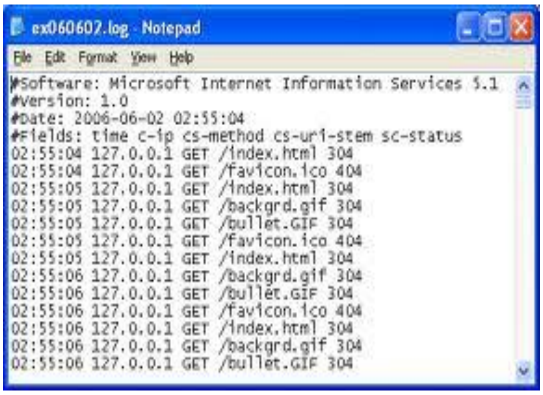

Users
User log files
Web server
User 1 log file
User 2 log file
User n log file
Identifying important key term from search factor using TF-IDF algorithm.
Removing noise from log files for obtaining access pattern.
Identifying specific user from above two factor

Fig.1. Architecture diagram of proposed system

All the log files were collected from web server for analyzing purpose. Fig.2. is the sample log files are shown which is collected from localhost server.



```
#Software: Microsoft Internet Information Services 5.1
#Version: 1.0
#Date: 2006-06-02 02:55:04
#Fields: time c-ip cs-method cs-uri-stem sc-status
02:55:04 127.0.0.1 GET /index.html 304
02:55:04 127.0.0.1 GET /favicon.ico 404
02:55:05 127.0.0.1 GET /index.html 304
02:55:05 127.0.0.1 GET /backgrd.gif 304
02:55:05 127.0.0.1 GET /bullet.GIF 304
02:55:05 127.0.0.1 GET /favicon.ico 404
02:55:05 127.0.0.1 GET /index.html 304
02:55:06 127.0.0.1 GET /backgrd.gif 304
02:55:06 127.0.0.1 GET /bullet.GIF 304
02:55:06 127.0.0.1 GET /favicon.ico 404
02:55:06 127.0.0.1 GET /index.html 304
02:55:06 127.0.0.1 GET /backgrd.gif 304
02:55:06 127.0.0.1 GET /bullet.GIF 304
```

Fig.2. Sample log file collected from local for a single session.

The given file has the factor in the following figure Fig.3. Usually these files consist of IP address, the time which the user have requested that page, the method the browser used to made that request. The Extended CLF contains these factors along with the browser version, referrer page and time factor.

In the notepad sample logfile, it consist of 4 factors about the user.

1. Time when user access the web page.
2. His system IP address (127.0.0.1)
3. The Method the browser used to get data (GET).
4. The web page the user requested.

These above criteria will be used as data mining factor.

Fig.3. Factors present in common log files

The three main stages of proposed system are data preprocessing, pattern discovery and pattern analysis. Data preprocessing involves removal of unnecessary data and gaining the necessary token using TF-IDF algorithm. Pattern discovery data mining techniques are used in order to extract patterns of usage from Web log data. The knowledge that can be discovered is represented in the form of rules and by combining both the data's from web logs and token obtained using TF-IDF algorithm for classification.

IV. EXTRACTION OF HIDDEN DATA FOR USER IDENTIFICATION

The task of user and session identification is to find out the different user sessions from the original web access log. User's identification is, to identify who access web site and which pages are accessed. The goal of session identification is to divide the page accesses of each user at a time into individual sessions. A session is a series of web pages user browse in a single access. The difficulties to accomplish this step are introduced by using proxy servers, e.g. different users may have same IP address in the log. ECLF is used to solve these problems in the proposed method. In the proposed system to distinguish a specific user from different users the IP address and their access pattern are used. Mostly single system may be used by many users, but their access pattern is different. In case the IP addresses may be same but when the request is made from different browsers and operation systems indicate different users. Sometimes same IP, same browser but different access pattern for same query also indicates that the user is different.

4.1 Identification of features from web logs:

Consider the following ECLF for token identification for token identification.
www.google.co.in - - [01/Jan/97:23:12:24 +0200] "GET /index.html HTTP/1.0" 200 1220 "https://www.google.co.in/webhp?ie=utf-8&oe=utf8&gws_rd=cr&ei=OlM7VYX1Loa6uATEmIDoAw#q=eclf+log+file "Mozilla/4.01 (X11; I; SunOS 5.3 sun4m)"
The ECLF features are shown in the tabulation Table.1. below.

Table.1. Known feature in ECLF

Field	Description	Value
%host	The name or IP of the remote host	*www.google.co.in*
%date	Date as Day:Hour or just Day	*01/Jan/97:23:12:24 or 1998-02-02*
%time	Hour	*23:12:24*
%hourshift	Shift from GMT	*+0200*
%method	Method requested to send the file	*GET*
%page	The file requested	*/index.html*
%protocol	Protocole used	*HTTP 1.0*
%status	The status code	*200*
%requetesize	The byte transfered for the requested file	*1220*
%agent	Browser and OS of the remote host	*Mozilla/4.01 (X11; I; SunOS 5.3 sun4m)*
%refer	The page the request come from	*http://www.google.co.in/*
%query	Arguments from the request	q=eclf+log+file

4.2 Identification of keywords using TF-IDF:

This process involved in removing unwanted words from known input file. Consider the same query in section 4.1, it contain necessary keyword as the domain name, referral page and the query and the unwanted words are the one which is not contributing anything in the process like "/, >, <. @, &, !, :, ;, is, was, or" etc.,. Term Frequency involves in identifying number of times a particular keyword is present. When applying same method to a ECLF for a particular session, it is identified that the term occur more frequently is identified as google, elclf, log, file, Mozilla. It is known that no two user have same search habit, therefore a specific user can be identified by combining both ECLF features along with the key term identified by TF-IDF algorithm.

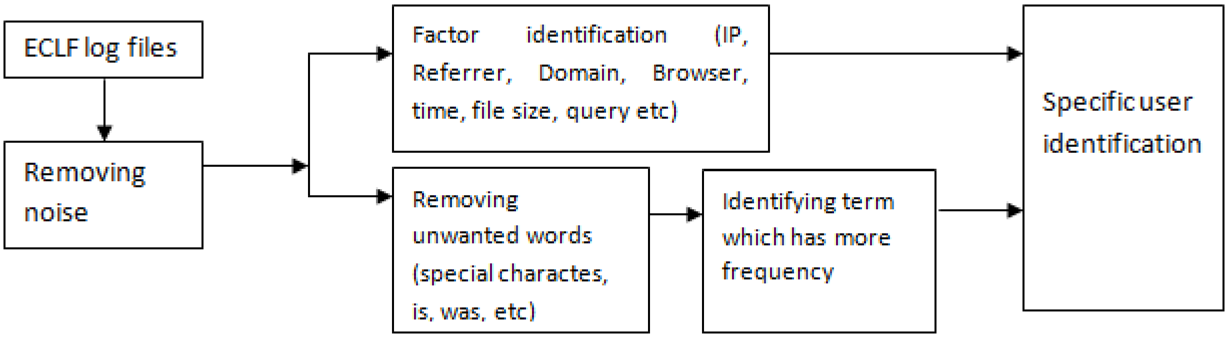

ECLF log files
Removing noise
Factor identification (IP, Referrer, Domain, Browser, time, file size, query etc)
Removing unwanted words (special charactes, is, was, etc)
Identifying term which has more frequency
Specific user identification

Fig.4. Identification of specific user from key term and log factors.

4.3 Performance analysis:

Both the log factors and the key term using TF-IDF algorithm can identify a specific user more precisely than the previous approach. To validate the effectiveness and efficiency of our methodology the proposed system uses web server logs at various period of time in the time span of one month. The pattern identified from the proposed system is as follows in table.2.

Table.2. Performance over time period

	Number of entries in web logs	**Number of Users**	**Number of unique user**
Jan	1020	320	112
Feb	1200	345	129
March	1345	375	133
April	1400	389	148
May	1650	411	152
June	1700	463	136
July	2000	521	162

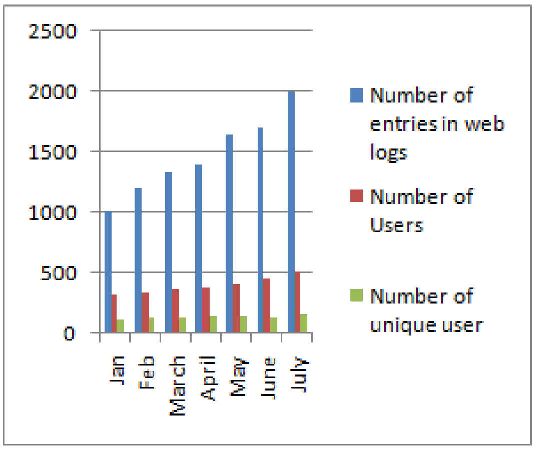


Fig.5. Result of Experiment

V. CONCLUSION AND FUTURE WORK

In this research the proposed system identify distinct user with accuracy of 92.3%. The proposed system combine both the key term identified using TF-IDF algorithm, the factor from ECLF logs to identify the distinct user. This technique is more useful for custom-making websites and to improve the design of WebPages. The result of the proposed system can be more accurate when it combined with other heuristic approaches like web structure mining and content mining.

REFERENCE

[1] Zhang Huiying, and Laing Wei, "An Intelligent Algorithm of Data Pre-processing in Web Usage Mining", Proceed-ings of the 5th world Congress on Intelligent Control and Automation, June15-19, 2004, Hangzhou, P.R.China.

[2] R. Kosala, and H. Blockeel, "Web Mining Research: A Survey", SIGKDD Explorations, 2(1):1-15, 2000.

[3] Hsinchun Chen, Xin Li, Michael Chau, Yi-Jen Ho, Chunju Tseng, Using Open Web APIs in Teaching Web Mining, ACM, 2009.

[4] Claudia Elena Dinuca, The process of data pre-processing for Web Usage Data Mining through a complete example, Annals of the "Ovidius" 2011.

[5] B. Mobasher, R. Cooley, and J. Srivastava. Creating adaptive web sites through usage-based clustering of urls. In IEEE Knowledge and Data Engineering Workshop (KDEX'99), 1999.

[6] C. R. Anderson. A machine learning approach to web personalization. PHD Thesis, University of Washington, 2002.

[7] Jian Pei, Jiawei Han, B. Mortazavi-asl & Hua Zhu: Mining Access Patterns Efficiently from Web Logs.

[8] N.S. Sudharsan, Dr. K. Latha. "Improvising Seeker Satisfaction in Cloud Community Portal: Dropbox". IEEE-International conference on Communication and Signal Processing, April 3-5, 2013, pp. 321-325.

[9] N.S. Sudharsan, Dr. K. Latha. "Preeminent System for Detecting Venomous Banking Sites in Online Business ". Applied Mechanics and Materials Vol. 573 (2014) pp 519-522.

[10] DeMin Dong, Exploring on Web Usage Mining and its Application , 5th world Congress on Intelligent Control and Automation, June 15-19,2004,China.

[11] Renata Ivancsy, and Sandor Juhasz, "Analysis of Web User Identification Methods". International Journal of Computer, Control, Quantum and Information Engineering Vol:1, No:10, 2007