A Survey on Web Usage Mining Techniques and Applications

Lya Hulliyyatus Suadaa
School of Electrical Engineering and Informatics
Bandung Institute of Technology
lya@stis.ac.id, lya@students.itb.ac.id

***Abstract*—Nowadays, Web becomes an important part of organization. The huge usage data as a result of interaction of users and Web can be extracted to be knowledge applied in various application. The main problem of Web usage mining are the nature of data they deal with and the handling methods. A survey on Web usage mining was conducted with Systematic Literature Review method to identify relevant studies about data sources, techniques, applications, and current issues that would be the key of future research direction in this area.**

Keywords—web usage mining; web log mining; web mining; data mining

I. Introduction

Nowadays, Web is a necessity of organization to be able to compete in the information age. A huge information not only can be extracted from the content of the Web, but also from communication of users and the Web. Web usage mining refers to the automatic patterns discovery and analysis in clickstreams, user transactions and other associated data collected or generated as a result of user interactions with Web resources on one or more Web sites [1]. It has been a potential technology for understanding behavior of the user on the Web[2].

The main problem of Web usage mining is the nature of data they deal with. As there are a lot of transaction between user and the Web by seconds, the volume of data which are not completely structured need to be extracted. Techniques to deal with this problem widely discussed to improve information quality. The extracted knowledge can be used in various application. The current of Web usage mining area would be described in this paper, completely with the issues as a key future research direction.

II. Methodology

The research method used in this study is a systematic literature review. A systematic literature review is an approach to identify, evaluate and interpret all relevant studies regarding a particular research question, topic area or phenomenon of interest [3]. The method is a form of secondary study.

A. Research Question

The research questions (RQ) of this paper are listed as follows:

RQ1. What are the data sources used in Web usage mining?

RQ2. What are the methods used to extract the knowledge?

RQ3. What are the applications of Web usage mining?

RQ4. What are the current issues in Web usage mining area?

B. Search Strategy

This study used online sources searched to collect the articles. The databases and the URLs are shown in table 1.

TABLE I. Online Sources Searched for Relevant Studies

Database	URL
IEEE Explore	http://ieeexplore.ieee.org/
Elsevier	http://sciencedirect.com/
ProQuest	http://search.proquest.com/

Strategy to search the articles were developed by using keywords combined Boolean operator.

> "web usage mining" OR "web log mining" OR "web log analysis"

Figure 1. Search String

978-1-4799-6527-4/14/$31.00 ©2014 IEEE

Inclusion and exclusion criteria were implemented in the selection process. The inclusion criteria are articles published in international journal between 2009 and 2013 that focus on Web usage mining application and use some methods to extract the knowledge from web usage. And the exclusion criteria are survey paper on web usage mining.

C. *Data Collection*

The data item collected from each paper were described in following table.

TABLE II. DATA COLLECTION FORM

Item ID	Field	Concern
F1	Author(s)	Documentation
F2	Year	Documentation
F3	Title	Documentation
F4	Keywords	Documentation
F5	Data Sources	RQ1
F6	Extraction Knowledge Methods	RQ2
F7	Web Usage Mining Applications	RQ3
F8	Future Work	RQ4

D. *Data Analysis*

In this section, we analyze the total number of papers to answer the research questions. In total, 324 papers were considered for the study derived from digital databases. From this set, 37 papers were included after applying exclusion criteria. Figure 2 shows article keyword clouds representing most popular keywords in web usage mining.

Figure 2. Article Keyword Clouds

The number of papers published is seen to have been increasing yearly, except for 2011. This shows that research on web usage mining is very active and still going on.

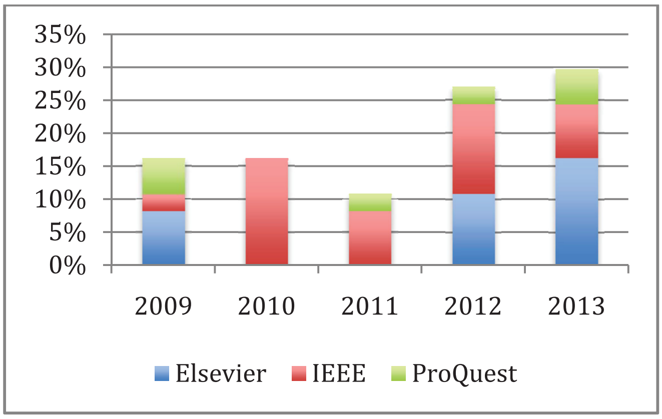

Figure 3. The Percentage of Publication per Year

III. RESULTS

There are three main process in Web usage mining that answer RQ1 to RQ3:

A. *Web Usage Mining Data Source*

Web usage mining data source located in three different location [4]:

- Server

  Web server log files provide most accurate and complete usage data. Most web usage mining techniques used this data.

- Client

  Client side data can be collected by remote agent or modified web browser. Client data collection depends on user cooperation either in enabling remote agent functionality or willingness to use modified browser [5].

- Proxy

  Proxy server is an intermediate level of caching between client browsers and web servers. Proxy server logs can be used to discover the usage pattern of a group of users, who share a common proxy server [6].

Table III shows the distribution of data source used in collected papers. Web server logs are used in almost web usage mining methods. There are still few researchers who combine the data from client and proxy side because of the difficulty of data collection process.

TABLE III. DATA SOURCES USED IN WEB USAGE MINING

Data Source	Percentage
Server	86%
Client	5%
Proxy	9%

B. *Web Usage Mining Techniques*

There are three main process in web usage mining :

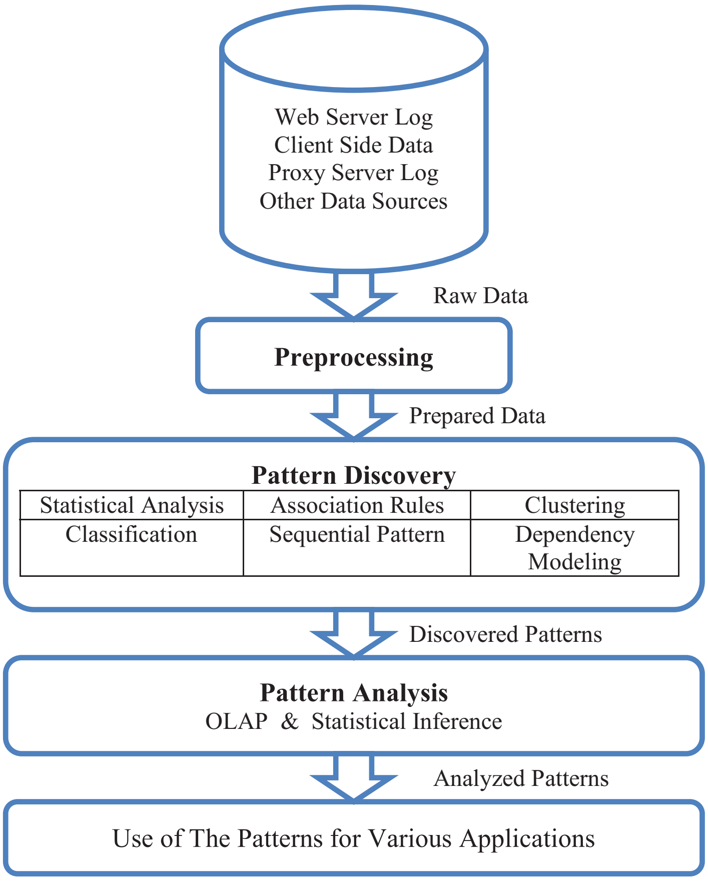

Figure 4. Web Usage Mining Process

- Preprocessing

  Preprocessing is a series of Web usage processing covering data cleaning, user identification, session identification, path completion and transaction identification [7]. This step is an important process before mining because the usage data are usually noisy and ambiguous.

- Pattern Discovery

There are methods used to extract the knowledge from web usage, commonly called pattern discovery, consist of:

➢ Statistical Analysis

Statistical analysis used by most web traffic analysis tools, especially descriptive statistical analysis with various tables and graphs. R.Mahajan, J.S. Sodhi, and V. Mahajan used this methods to produce useful information to build adaptive e-learning site [8].

➢ Association Rules

In web usage mining, association rules refer to sets of pages that are accessed together with a support value exceeding some specified threshold. The methods are usually combined with clustering to find interesting relation in each cluster resulted. [9][10][11]

➢ Clustering

Clustering is the process of partitioning a set of data into subset according to their similarity [12]. There are two types of clustering that interesting to do: users clustering and pages clustering. B.N. Devi, Y.R. Devi, B.P. Rani and R.R. Rao applied users clustering to discover user clusters with similar behavior pattern and preferences [13]. It is also can be used in security area to detect malicious and non malicious website visitors [14]. On the other hand, pages clustering find page groups with similar page contents.

➢ Classification

Classification is the techniques learning a classification function from data that are labeled with pre-defined classes or categories [1]. Classification techniques usually applied to build the users model according to various predefined metrics. For example, in e-commerce area, given a set of user transaction, classification model can be built to classify users into those who have a high propensity to buy and those who do not.

➢ Sequential Pattern

Sequential pattern record the navigational behavior of the user and stochastic techniques that use the sequence of Web pages that have been visited in order to predict subsequent visits. B. Verma, K. Gupta, S. Panchal, and R. Nigam used sequential pattern mining to discover extracted patterns which are used to generate page recommendations at run time [15].

- Dependency Modeling

Dependency modeling is the process to build a model to predict various variables. V.V.R.M. Rao and V.V. Kumari introduces an efficient hybrid predictive model, which is a combination of Markov model and Bayesian theorem in predicting the web pages visiting by users [16].

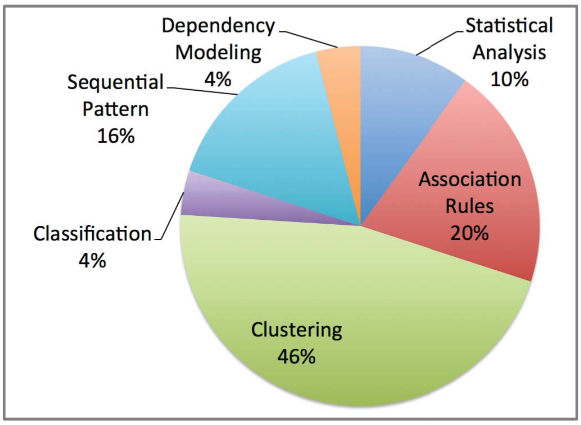

Figure 5. Knowledge Extraction Methods

- Pattern Analysis

The pattern discovered in pattern discovery process are analyzed in this step. The most common form of pattern analysis are On-Line Analytical Processing (OLAP) and statistical inference methods.

C. *Web Usage Mining Applications*

There are various current Web usage mining applications. Figure 4 shows proportion of web usage mining application in collected papers. Recommendation system and site modification are the most popular application in this area.

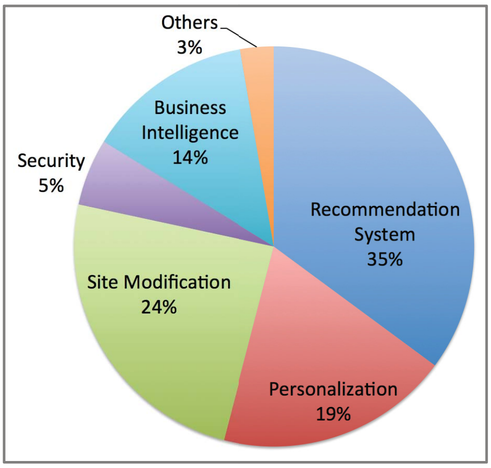

Figure 6. Web Usage Mining Applications

- Recommendation System

Recommendation systems are widely used on the Web for recommending products and services to users. This system mostly used in e-commerce and e-learning web.

- Personalization

Personalization is the process of tailoring pages to individual users characteristics or preferences. Web usage has been used to model the users' interests and personalize Web applications [17].

- Site Modification

Web usage provides detailed feedback on user behavior that can be useful for website designer to improve usability of web.

- Security

Web usage mining is also applied in security area. R. Tamimi and M.E. Mohammadpourzarandi used web mining techniques for detecting, preventing and predicting cyber attacks on virtual space [18].

- Business Intelligence

Knowledge of interaction between users and web can be used to discover marketing intelligence and improve customer relationship management.

D. *Web Usage Mining Current Issues*

There are some issues discussed in web usage mining area. Web usage combined by other resource can enrich data quality used in usage mining processes. For example combination of web content and web usage can improve pattern quality [19]. In methods side, combination of some methods can extract more knowledge, including techniques that support a very high dimensional data space with huge amount of data. The application in cloud computing area are also interesting to be discussed. On the other hand, privacy issue is a sensitive topic for most users who want the anonymity on the web.

IV. CONCLUSSION AND FUTURE WORK

Web usage mining area is interesting to be studied as a result of a huge usage data today. Data sources of Web usage are mostly from Web server logs that could be combined with client side data and proxy logs. Data mining techniques applied to extract the knowledge are statistical analysis, association rules, clustering, classification, sequential pattern and dependency modeling. The extracted knowledge are very useful in various applications consist of recommendation system, personalization, site modification, security and business intelligence.

Information quality of web usage is the main focus of the future works. The quality can be improved from each step of Web usage mining process. In preprocessing step, web usage combined by other resources to enrich the data. In pattern discovery and analysis step, hybrid methods were proposed to mine more knowledge. Moreover, it is interesting to discuss the application in new technology area like cloud computing. On the other hand, there are privacy issue which is a sensitive topic for most users who want the anonymity on the Web.

REFERENCES

[1] B. Liu, Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data, 2nd ed., New York: Springer, 2011.

[2] P. Nithya and P. Sumathi, "A survey on Web usage mining: theory and applications," International Journal Computer Technology and Applications, Vol 3, 2012, pp. 1625-1629.

[3] Kitchenham B., S. Charter, "Guidelines for performing systematic literature reviews in software sngineering," Version 2.3, EBSE Technical Report EBSE-2007-01, Keele University and University of Durham, UK, 2007.

[4] S. Langhnoja, M. Barot, and D. Mehta. "Pre-processing: procedure on web log file for web usage mining", International Journal of Emerging Technology and Advanced Engineering, Vol 2, 2012.

[5] J. Srivastava, R. Cooley, M. Deshpande, P. Tan. "Web usage mining: discovery and applications of usage patterns from web data", ACM SIGKDD, Vol 1, 2000.

[6] S. Dhawan and M. Lathwal, "Study of preprocessing methods in web server logs", International Journal of Advanced Research in Computer Science and Software Engineering, Vol 3, 2013.

[7] V. Chitraa and A.S. Davamani, "A study on preprocesssing methods for web usage data", International Journal of Computer Science and Information Security, Vol 7, 2010.

[8] R. Mahajan, J.S. Sodhi, V. Mahajan, "Web usage mining for building an adaptive e-learning site: a case study", International Journal of e-Education, e-Business, e-Management and e-Learning, 2014.

[9] C.J. Carmona, S. Ramírez-Gallego, F. Torres, E. Bernal, M.J. del Jesus, S. García, "Web usage mining to improve the design of an e-commerce website: OrOliveSur.com", International Journal of Expert System with Applications, 2012.

[10] Y. Slimani, A. Moussaoui,Y. Lechevallier, A. Drif, "A community detection algorithm for Web Usage Mining Systems", International Symposium on Innovation in Information & Communication Technology, 2011.

[11] J.D. Vela´squez, "Combining eye-tracking technologies with web usage mining for identifying Website Keyobjects", International Journal of Engineering Applications of Artificial Intelligence, 2013.

[12] J. Han, M. Kamber, and J. Pei, Data Mining: Concepts and Techniques, Morgan Kaufmann Publisher, 2011

[13] B.N. Devi, Y.R. Devi, B.P. Rani, R.R. Rao, "Design and Implementation of Web Usage Mining Intelligent System in the Field of e-commerce", International Conference on Communication Technology and System Design, 2012.

[14] D. Stevanovic, N. Vlajic, A. An, "Detection of malicious and non-malicious website visitors using unsupervised neural network learning", International Journal of Applied Soft Computing, 2013.

[15] B. Verma, K. Gupta, S. Panchal, R. Nigam, "Single Level Algorithm: An Improved Approach for Extracting User Navigational Patterns To Improve Website Effectiveness", International Conf. on Computer & Communication Technology, 2010.

[16] V.V.R. M. Rao and V. V. Kumari, "An Efficient Hybrid Predictive Model to Analyze the Visiting Characteristics of Web User using Web Usage Mining", International Conference on Advances in Recent Technologies in Communication and Computing, 2010.

[17] G. Paliouras, "Discovery of Web user communities and their role in personalization", User Model User-Adap Inter, 2012.

[18] R. Tamimi, M. E. Mohammadpourzarandi, "The Application of Web Usage Mining In E-commerce Security", International Conference on e-commerce in developing countries with focus on e-security, 2013.