Volume 3, Issue 4, April-2016, pp. 204-209 ISSN (O): 2349-7084

International Journal of Computer Engineering In Research Trends

Available online at: www.ijcert.org

A Survey on Web Page Recommendation and Data Preprocessing

[1]Ms. Sonule Prashika Abasaheb, [2]Prof. Tanveer I. Bagban

1 (M.E. PART-II, Department of Computer Science and Engineering, D.K.T.E. Society's Textile and Engineering Institute, Ichalkaranji, Shivaji University, Kolhapur, Maharashtra, India. Email: prashikasonule@gmail.com)

2 (Associate Professor, Department of Information Technology, D.K.T.E. Society's Textile and Engineering Institute, Ichalkaranji, Shivaji University, Kolhapur, Maharashtra, India. Email: tbagban@yahoo.com)

Abstract: - In today's era, as we all know internet technologies are growing rapidly. Along with this, instantly, Web page recommendations are also improving. The aim of a Web page recommender system is to predict the Web page or pages, which will be visited from a given Web-page of a website. Data preprocessing is one basic and essential part of Web page recommendation. Data preprocessing consists of cleanup and constructing data to organize for extracting pattern. In this paper, we discuss and focus on Web page Recommendation and role of data preprocessing in Web page recommendation, considering how data preprocessing is related to Web page recommendation.

Keywords - Recommender System, Web server logs, Web mining, Web usage mining, Data Preprocessing.

———————————— ◆ ————————————

I.INTRODUCTION

The unpredictable increase and growth of information on the World Wide Web, with the progress of innovative electronic devices, has made information of Web increasingly important in everyone's life. In today's era, as we all know internet technologies are growing rapidly, so web has become large storage of information and this amount of information grows with high and rapid rate of change without any control of editor; consequently, websites are also introduced rapidly with new innovations.

Web page recommender systems can recommend Web pages automatically which are most interesting to a particular user based on that current Web navigation behaviour of user. Regularly, Web users have to struggle for finding useful pages and are very probable to leave the site, if the index pages of a website are not well designed. Thus, even though recommenders have been appreciated and valued in experimental research, they have not been commercially successful [1]. These are one of the problems faced by user while accessing most interested Web-pages on website.

The main reason behind these above given problems, is the huge amount of explosive growth of information, which is irrelevant and noisy. Considering the above mentioned problems, it seems that there was a need of cleaning and construct or structure that irrelevant data. This constructing data from noisy and complex data, is nothing but data preprocessing, we will elaborate this in section 4.

The main problematic thing and difficulty is in accessing most interested Web pages. Problems like this relates with usage of Web. Hence, there is a need of cleaning and constructing or structuring Web log data, which is nothing but data preprocessing part in Web Usage Mining [3]. Data preprocessing plays a vigorous role because of redundant irrelevant log data nature [4]. Thus, we find that, data preprocessing is one basic and essential part of Web-page recommendation. This paper is structured as below: Section 2 comprises a review of Web page recommendation. Section 3 clarifies categorization of recommendation system and web mining, and it discusses how data preprocessing is related to Web page recommendation. Section 4 illustrates data preprocessing and its steps. Section 5 provides comparative analysis of data preprocessing techniques use; and finally, section 6 gives the conclusion.

II. WEB PAGE RECOMMENDATION

In today's era, along with the rapid growth of internet technologies, there is fast increase in innovated websites. This overpowers Web users, by offering many choices. Therefore, Web users sometimes, probably make poor decisions when they surf the Web. So there is a need of a system that recommends the Web pages to Web user, to make their work easy, such systems are nothing but Web page recommender system.

In 1996, the term recommender system was first invented at a workshop, and has been used inconsistently and imprecisely in published work [1]. Web-page recommender systems have become valuable increasingly for helping Web users to find the most interesting and important Web-pages on specific websites. Best Web page recommendations can improve website usage along with Web user satisfaction. The important characteristic of the recommendation system is to study from current user's historic data and also from remaining users. The recommendation system decides current user's domain from the historic data of user, then pages prediction is done according to the domain of user ([4], [8]).

III. TAXONOMY

The recommendation system is one of the applications exploited by results from web usage mining [4]. Thus, Web Page recommendations are mostly related to web mining. Therefore, we will see categorization of both Web Page recommendation Technique and Web mining in this section and how data preprocessing is related to Web page recommendations.

A. Classification of recommendation technique:

The recommenders are normally implemented by filtering algorithms categorized into three main types, depending on how the recommendations are performed. Thus, recommendation process or methods can be classified as below, according to source of knowledge used by them for recommendations ([5], [6]).

Content based recommendations: This type of systems use, Content-Based algorithms (CB), which filter, clean and recommend articles and items that are similar to others accessed by the user in the past.

Collaborative Recommendations: Such process uses, Collaborative Filtering (CF) algorithms, that clean, filter and recommend articles and items based on the other user preferences. For example, it recommends items which user has not accessed in the past but, mentors of that user liked and accessed more in the past.

Knowledge-based Recommendation: To generate a recommendation, this method use domain knowledge about users and items, thus need to clean users log data.

Hybrid Recommendations: This method combines methods or techniques of two or more recommendations from above explained categories, in order to gain better optimization of system.

B. Classification of Web mining

A common classification or taxonomy of Web mining is done into three different types; Web content mining, Web structure mining and Web usage mining ([2], [3], [4]).

1) Web Content Mining: This is the process of extraction and integration of useful information, data and knowledge from Web page contents ([3], [10]). It is the process to discover useful information and web page contents like video, audio, hyperlinks as well as metadata [2].

2) Web Structure Mining: A graph theory is used to analyse the node and connection of a website structure in this process. A web graph structure, contains web pages as nodes and hyperlinks as like edges connecting pages which are related to each other ([2], [3], [4]).

3) Web Usage Mining: This is the data mining technique application to discover usage patterns from web data. In our viewpoint, access logs on server side are the usage data which keeps user navigation information [3]. We will focus on Web Usage Mining in this section, as the recommendation system is one of applications exploited by results from web usage mining [4].

3.1) Web Usage Mining: Web Usage Mining is a portion of web mining which compacts and deals with the interesting knowledge extraction and abstraction from log files. Usage mining tools discover and predict user behavior, in order to help the designer to improve the web site, such that regular users can get a personalized and adaptive service or to attract visitors. Web Usage Mining is also termed as Web Usage Analysis or Web Log Mining or Analysis of Click Stream [3].

Web Usage Mining consists of three main phases. These three phases are Data preprocessing, Pattern discovery and Pattern analysis ([2], [3]). Among these three phases, data preprocessing plays a vigorous role because of redundant and noisy nature of log data. Therefore, in next section we will focus on data preprocessing as it is basic and essential part of Web-page recommendation. This unit gives an overview of these three phases as below.

Data preprocessing: It consists of cleanup and constructing data to organize for extracting pattern. The Data preprocessing phase includes some steps or techniques ([2], [3]). Several data preprocessing techniques have been used in improving other phases of web usage mining like Pattern discovery and Pattern analysis.

Pattern discovery: This phase deals with mining and extraction of information from preprocessed data i.e. results of data preprocessing phase. Techniques from different fields such as data mining, pattern recognition and machine learning are applied to web usage data in order to discover user's web access patterns [3].

Pattern analysis: This is the final stage of Web Usage Mining. Its goal is to eliminate the irrelative patterns in order to extract the user interesting patterns from the results of the pattern discovery process explained above [3]. Figure 1, below shows these Web usage mining phases, by a basic structure [3].

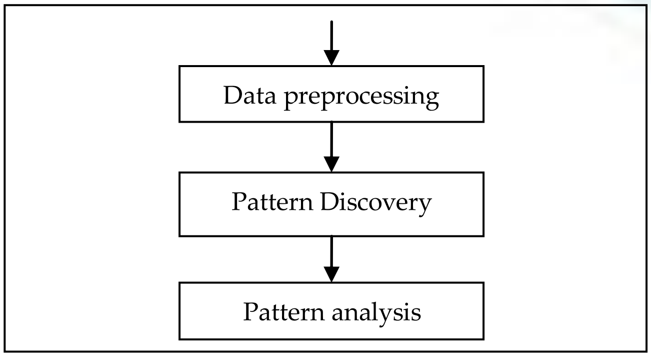

Figure 1. Basic structure Web Usage Mining.

IV. DATA PREPROCESSING

According to recommendation technique categories, each of it require to clean and filter data. Data preprocessing converts data into a format which will be more efficiently and easily managed user's purpose. The main data preprocessing task is to choose standardized data from the initial log files, organized for algorithm of user navigation pattern discovery [7]. This section discusses the significance of data preprocessing methods and steps involved in getting required content, effectively.

A] Data preprocessing steps:

Data preprocessing steps are methods or techniques that can be applied according to source file available. These preprocessing techniques are different for different sources of log files. Various authors have shown that which log file source needs which preprocessing technique [2]. We will see this in section V below.

Various steps involved in data preprocessing phase are shown through the figure 2; below, as given in [3]:

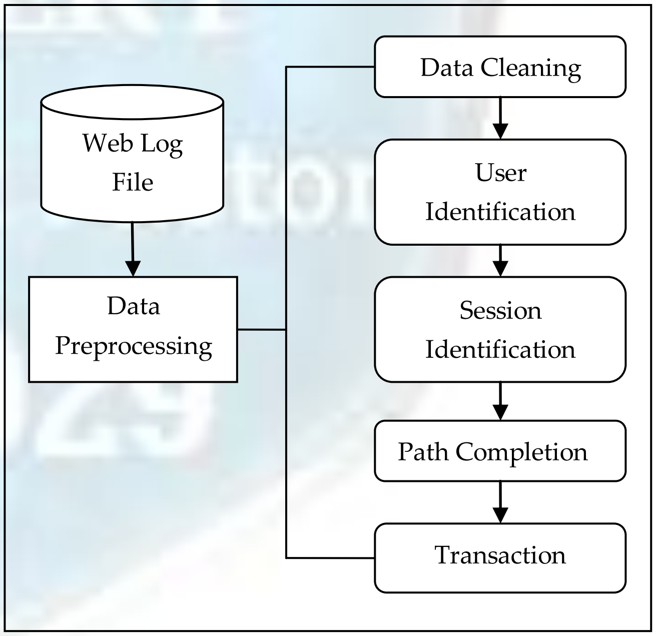

Figure 2. Data preprocessing steps.

1) Data Cleaning: It is a process of removing items such as gif files, jpeg or sound files and references due to spider navigations which are irrelevant. Data quality which will be improved also improves analysis on data quality ([3], [8]).

2) User Identification: This is an important step in web usage mining of individual user identification that accesses a web site. This is essential phase to determine who has accessed Website and which pages are mostly accessed ([3], [8]).

3) Session Identification: A sequence or series of web pages that user browses in a single access is called as session. The objective of session identification is to discover and divide the page accesses of each user session, into individual separate sessions ([3], [8], [9]).

4) Path Completion: Path completion is used to obtain the complete user access path. This step is difficult but important as this is final step in which user session file is completed [3].

5) Transaction Identification: The aim of transaction identification is to create important reference cluster for each user. So, this is done by merging or dividing approaches. Both approaches has a transaction list and some parameters as an input; and a final transaction list to be operated on by a function in the module which is same like input format, is nothing but the output [3].

V. COMPARATIVE ANALYSIS

As shown in section two Web page recommendation systems, use different algorithms according to the type of recommender system. The algorithms like Content-Based algorithms (CB) and Collaborative Filtering (CF) algorithms are used by recommendation system based on their type. Content based recommendations systems use, Content-Based algorithms (CB). Collaborative Recommendations systems use, Collaborative Filtering (CF) algorithms, that clean, filter and recommend articles and items based on the other user preferences ([5], [6]).

Data Preprocessing Techniques and algorithms can be applied according to source file available. These preprocessing techniques explained above are different for different sources of log files. Various Authors has shown that which log file source needs which preprocessing technique and which algorithms can be applied.

For example, (log file source) Server Log File needs preprocessing techniques Data Cleaning, Log File Filtering, Session Identification and User Identification; and Source log file like- English Study Web site Log File needs techniques such as Data Cleaning, User Identification, Session Identification, Path Completion, Transaction Identification; and algorithms, like 'Reference Length' and 'Maximal Reference Length'. This is illustrated by some examples in analysis table I as mentioned below [2].

TABLE I. ANALYSIS FOR PREPROCESSING BASED ON SOURCE OF LOG FILE

Source of log file	**Preprocessing technique**	**Algorithms Applied**	**Authors**
English Study Web site Log File	Data Cleaning User Identification Session Identification Path Completion Transaction Identification	Maximal Forward References(MFR), Reference Length	Yan LI, Boqin FENG and Qinjiao MAO [10]
IIS Server Log File	Data Cleaning User Identification Session Identification Path Completion	Based on referred web page and fixed priori threshold	Ling Zheng, Hui Gui and Feng Li [11]
Web server Log file	Data Preprocessing	Based on Collaborative Filtering	JING Chang-bin and Chen Li [12]
Chizhou College Website	Data Filtering Session Identification	Frame page and Page Threshold	Fang Yuankang and Huang Zhiqiu [13]

According to the table above, the details of algorithms and data preprocessing implementation of web usage mining are presented by Yan Li's paper [10]. It explains the reference length algorithm, which modifies the reference length of pages in complete path by considering the average reference length of auxiliary pages which is estimated and valued in advance through the maximal forward references and reference length algorithms.

FP-Growth Algorithm was used by Huiping Peng [14] for the web log records to be processed and a set of frequent access patterns to be accessed. Then using both combinations of site topology interestingness and browse interestingness of association rules for web mining a new pattern to provide valuable data for the site construction was exposed.

To solve some problems that exist in traditional data preprocessing technology for web log mining, an improved data preprocessing technology is used by the author ling Zheng [11]. The algorithms based on 'fixed priori threshold' and 'referred web page' were used. These were applied on IIS Server Log file source of log file. Data Cleaning, User Identification, Session Identification and Path Completion were the preprocessing techniques applied on such sources of file.

A Web log data preprocessing algorithm, which is based on collaborative filtering, was brought by JIANG Chang-bin and Chen Li [12]. Even though statistic data are not enough and records visiting user history are absent it can perform user session identification fast, rapid and flexibly.

Algorithms named as Frame page, Page Threshold were applied on Chizhou College Website. Data preprocessing techniques; Data Filtering and Session identification were used as in [13].

VI. CONLUSION

This paper illustrated Web page recommendation and its types to know what Web page recommendation is and later it focuses on how data preprocessing is important part of Web page recommendation. Thus, data preprocessing plays a vigorous role in reducing and removing redundant, noisy and irrelevant nature of log data; and it is basic phase and essential for Web page recommendation.

REFERENCES

1. Ben Schafer, Joseph A. Konstan, and John T. Riedl, "Recommender Systems for the Web".

2. Vijayashri Losarwar et al., "Data Preprocessing in Web Usage Mining" International Conference on Artificial Intelligence and Embedded Systems July 15-16, 2012 Singapore.

3. Naga Lakshmi et al., "An Overview of Preprocessing on Web Log Data for Web Usage Analysis", International Journal of Innovative Technology and Exploring Engineering ISSN: 2278-3075, Volume-2, Issue-4, March 2013.

4. Mitali Srivastava, Rakhi Garg, "Preprocessing Techniques in Web Usage Mining: A Survey".

5. J. Manuel Adán-Coello, C. M. Tobar, Y. Yuming, "Improving the Performance of Web Service Recommenders Using Semantic Similarity". In JCS&T Vol. 14, No. 2, October 2014.

6. Sabanaz S. Peerzade, Vanita D. Jadhav, "A Review on Web Service Recommendation System Using Collaborative Filtering", Volume 3, Issue 3, March 2015.

7. Chaoyang Xiang, Shenghui He and Lei Chen, "A Studying System Based On Web Mining", IEEE International Symposium On Intelligent Ubiquitous Computing and Education, pp.433-435, 2009.

8. "A Survey on Preprocessing Methods for Web Usage Data", (IJCSIS) International Journal of Computer Science and Information Security, Vol. 7, No. 3, 2010.

9. R. Cooley, B. Mobasher, J. Srivastav (1999), "Data preparation for mining world wide web browsing pattern" in Journal of Knowledge and Data Engineering Workshop, IEEE, Vol.1 .

10. Yan LI, Boqin FENG and Qinjiao MAO, "Research on Path Completion Technique in Web Usage Mining", IEEE International Symposium on Computer Science and Computational Technology, pp. 554-559, 2008.

11. Ling Zheng, Hui Gui and Feng Li, "Optimized Data Preprocessing Technology For Web Log Mining", IEEE International Conference on Computer Design and Applications, pp. VI-19-VI-21, 2010.

12. JING Chang-bin and Chen Li, "Web Log Data Preprocessing Based on Collaborative Filtering", IEEE 2nd International Workshop on Education Technology and Computer Science, pp.118-121, 2010.

13. Fang Yuankang et al., "A Session Identification Algorithm Based on Frame Page and Page threshold", IEEE Conference, pp.645- 647, 2010.

14. Huiping Peng, "Discovery of Interesting Association Rules Based on Web Usage Mining", IEEE Conference, pp.272-275, 2010.

AUTHOR PROFILE

1. Ms. Sonule Prashika Abasaheb is student of M.E. PART-II, Department of Computer Science and Engineering, D.K.T.E. Society's Textile and Engineering Institute, Ichalkaranji, Shivaji University, Kolhapur, Maharashtra, India. Her interest area is Web Mining.

2. Prof. Tanveer I. Bagban is working as associate professor in, department of Information Technology, D.K.T.E. Society's Textile and Engineering Institute, Ichalkaranji, Shivaji University, Kolhapur, Maharashtra, India. He has 14 years of Teaching Experience. His interest area is Web Mining, Information Extraction. He has published about 4 papers in international journals and 2 papers in National Conferences.