A Survey On Various Approaches For Webpage Recommendation System In Web Mining

Dr. Shyamal Tanna, Darshan K Prajapati
Department of Computer Engineering
L.J. Institute Of Engineering and Technology, Ahmedabad, India

***Abstract*— Web mining is the application of data mining on web data and web usage mining is an important component of web mining. The goal of web usage mining is to understand the behavior of web site users through the process of data mining of web access data. Knowledge obtained from web usage mining can be used to enhance web design, introduce personalization service and facilitate more effective browsing. Webpage Recommender systems are intelligent systems which make suggestions about user's next webpage. Many of the largest Web sites are already using recommender systems to help their customers find a next webpage. Recommender system has become an important part of any entertainment or E-Commerce website. Various personal services in business play important roles in the success of current marketing field. This paper presents a review of literature containing latest works done in this field.**

***Index Terms*— Recommendation, Webpage Recommendation, Clickstream Data, Pattern Recongnizatioin, Web Usage Mining**

I.INTRODUCTION

The number of Internet applications has grown and continue to grow significantly, affecting the lives of people in various aspects of their life including education, health, business and etc. The convenience and flexibility of services offered by web applications are the contributing factors why web applications are fast gaining popularity. In the process, web applications almost invariably churn out huge data containing user transactions and activity logs of user operations. Within the broad conceptual framework of Knowledge Discovery from Databases (KDD), many studies have been conducted to explore ways of extracting potentially useful information embedded from large databases which can enhance decision making process. The core process of KDD, referred to as data mining, constitutes a number of different tasks aimed at extracting frequent patterns including association rules and sequential patterns mining. The application of data mining on web data is termed as web mining [6].

Rosli Omar, Abu Osman Md Tap, Zainatul Shima Abdullah [6] have further categorized web mining into three main components: web usage mining(WUM), web structure mining(WSM) and web content mining(WCM). WCM is the task of discovering useful information based on the content of web pages. Web contents include multimedia data, structured content such as XML documents, semi-structured such as HTML documents and unstructured data such plain text. Web content mining applications include the task of organizing and clustering the web pages based on content and as well as ranking of web pages based on contents. WSM focuses on the structure of web sites using source data in the form of the structural information present in Web pages; typical applications are link-based categorization of Web pages, ranking of Web pages through a combination of content and structure and reverse engineering of Web site models.

1.1 Web Mining

It is the application of data mining techniques to discover patterns from the World Wide Web. Web mining has mainly three types.

• **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 of Web users along with their browsing behavior at a Web site.

• **Web Structure Mining:** Web structure mining is the process of using graph theory to analyze the node and connection structure of a web site. According to the type of web structural data, web structure mining can be divided into two kinds: first Extracting patterns from hyperlinks in the web And second is Mining the document structure.

• **Web Content Mining:** Web content mining is the mining, extraction and integration of useful data, information and knowledge from Web page content. The heterogeneity and the lack of structure that permits much of the ever-expanding information sources on the World Wide Web, such as hypertext documents, makes automated discovery, organization.

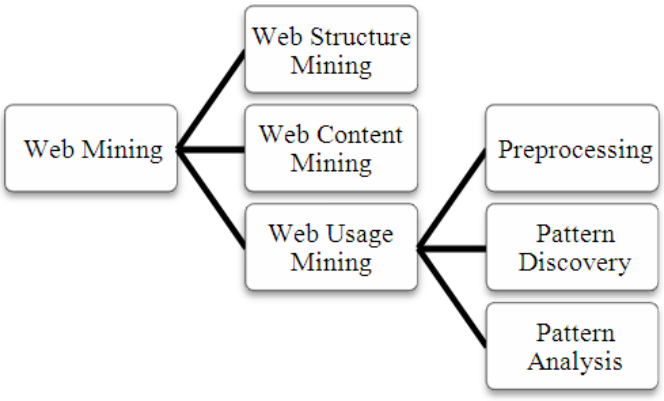

Fig1. Taxonomy Of Web Usage Mining

1.1.1 Collaborative filtering

These approaches building a model from a user's past behavior (items previously purchased or selected and/or numerical ratings given to those items) as well as similar decisions made by other users. This model is then used to predict items (or ratings for items) that the user may have an interest in. Collaborative filtering approaches often suffer from three problems: cold start, scalability, and sparsely.[7]

• There are two major approaches for collaborative filtering algorithms:

1. **Model-based approaches:** These approaches use training data to generate a model. These models have been used to predict the ratings for the items that a user has not been rated before. In this approach the raw data is usually processed offline. For example, decision trees, aspect models, latent factor models and clustering methods are model-based approaches for collaborative filtering.

2. **Memory-based approaches:** These approaches look at similar users or items based on their previous rating and combine their ratings in order to make new predictions. In this approach, the raw data is kept and processed in memory. Examples of memory-based collaborative filtering algorithms are user-based and item based methods.

• Item-based CF :

This method tries to predict users opinion on different items and then recommends an item based on the users transaction history as well as a numerical value that expresses the predicted likeliness of an item about which the user has not given his opinion.

• User-based CF [8]:

It certain majority of the customer have same taste then they join into the one group. Recommendation is given to user based on evolution of items by other users from the same group.

1.1.2 Content-based filtering

This approach uses a set of discrete characteristics of an item to recommend more items with similar properties. It is based on item description and user preference. For example, the music recommendation 'Pandora' uses the properties of a song or the singer in order to create a station which plays music with similar properties. In content based recommendation you compare items based on their features for movies things like title, genre, release date, director, producers, studio, etc.

1.1.3 Demographic

This Recommendation provides recommendations based on a demographic profile of the user. Recommended products can be produced for different demographic slots, by combining the ratings of users in those slots.

1.1.4 Knowledge-based

It suggests products based on inferences about a user's needs and preferences. This knowledge will sometimes contain explicit functional knowledge about how certain product features meet user needs.

II. BACKGROUND TERMINOLOGIES

This section describes recent development in the area of Web Usage Mining(WUM), from the perspective of the different types of patterns being mined.

A. Basic Mining Algorithms

At the heart of WUM are the generic mining algorithms which perform the task of extracting frequent patterns from data files. Some of these algorithms include AprioriAll[10], Generalized Sequential Pattern (GSP) [10], Sequential Pattern Discovery using Equivalence classes (SPADE), Frequent Pattern-Projected Sequential Pattern mining (FreeSpan) and Prefix-Span [11]. Although these algorithms have the same objective of mining for frequent patterns, they employ different methods to achieve the goal. For WUM specific purposes, the algorithms are required to process only single-element sequences which are suitable for web navigational sequences.

The Apriori-based algorithms require multiple scans of database. For each iteration i, the algorithms generate candidate itemsets of size i by joining frequent itemset of length with itself and subsequently pruning out candidates which contain infrequent subsequences. By Apriori property, these candidates cannot be frequent since all subsets a frequent itemset are frequent. To compute the support values for all candidate itemsets being generated, the database is scanned again. Candidate itemsets which satisfy support threshold are considered frequent itemsets. The process

continues until all frequent itemsets are discovered. This means that the database is access at least k number of times, where k is the maximum number of iterations it takes to mine all frequent itemsets. This is one of the main drawbacks of Apriori method.

B. Mining association rules from WUM

Mining association rules from web data is well studied due to its popularity. Association rules in WUM describe the relationships between two or more web pages. For example, the association rule pattern page1 page3, where page1 and page3 are pages within a set of pages contain in user access session, states that sessions that contain page1 will most likely also contains page3. In other words, association rules in WUM describe web pages which are frequently visited together. Having this knowledge may help a web designer restructure the website by placing frequently visited pages closed to each other so as to enhance the speed of browsing. Web usage mining includes three phases namely pre-processing, pattern discovery and pattern analysis. Their method combines the technique of clustering and association rule mining to extract significant user behaviors.

C. Mining for Classification and Clustering

An application of WUM for predicting user traversal pattern for a colleges web site can be found in [26]. This paper describes web usage mining for the college log files to analyze the behavioral patterns and profiles of users interacting with a Web site. The discovered patterns are represented as clusters that are frequently accessed by groups of visitors with common interests. In this paper, the visitors and hits were forecasted to predict the further access statistics. In the e-commercial recommendation system, users are clustered according to features like the browsing time and frequency[9].

D. Mining Sequential Patterns in WUM

The application of sequential mining techniques on WUM from web log accesses, involves mining for sequential patterns of web pages visited by web users in order to understand user browsing behavior. Web access pattern tree (WAP-tree) mining [20] is a sequential pattern mining technique for web log access sequences. In the first stage, it scans the sequence database while constructing a compact prefix tree to store the web access sequence database. In the second stage, the mining algorithm uses the WAP-tree to mine frequent sequences from the WAP-tree. The mining process involves recursively re-constructing intermediate trees, starting with suffix sequences and ending with prefix sequences. The strength of WAP-tree method includes avoiding the time consuming candidate generating process. However, WAP-tree algorithm requires recursive reconstruction of intermediate WAP-trees during mining process. This process is very time-consuming.

III. LITERATURE REVIEW

Mr.M.Saravanan and,Dr.V.L.Jyothi [1] proposed the optimal sequence of pages in log file. Log files has huge amount of data so to obtain sequential pages first data preprocessing is need to be done and then genetic algorithm has to be applied to get the optimal sequence of pages. Genetic algorithm helps in large amount of data input. The sequential pages of visited user is presented which enhances the pre-processing steps of web log usage data in data mining. Firstly, user identification and session identification and session identification is applied on the log files. This pre-processing step filters the number of users and the number of unique users. Sequential access table is the combination of both user and session identification. All the sequential pages visited by the users can be seen in this table. For getting the best sequential pages, genetic algorithm is used. And lastly best sequential pages is obtained.

R.Rathipriya, Dr. K.Thangavel, J.Bagyamani[2] proposed a bi-clustering approach for web data, which identifies groups of related web users and pages using spectral clustering method on both row and column dimensions. biclustering algorithms are widely applied to the gene expression data. Most of these algorithms are failed to extract the coherent pattern from the data matrix. In web mining, there is no related work that has been applied specific biclustering algorithms for discovering the coherent browsing patterns. In this paper, Greedy Search Procedure and evolutionary approach namely Genetic Algorithm (GA) is introduced to obtain the optimal coherent browsing patterns. The results show that GA outperforms the greedy procedure by identifying coherent browsing patterns. These patterns are very useful in the decision making for target marketing.

Hiral Y. Modi , Meera Narvekar [3] proposed an Online recommendation System. The system involved two phases that work in conjunction with each other i.e. the online and offline phase. Data pretreatment and navigation pattern mining is carried out in offline phase while predictions are generated in the online phase. They also proposed the online and offline phase architecture.

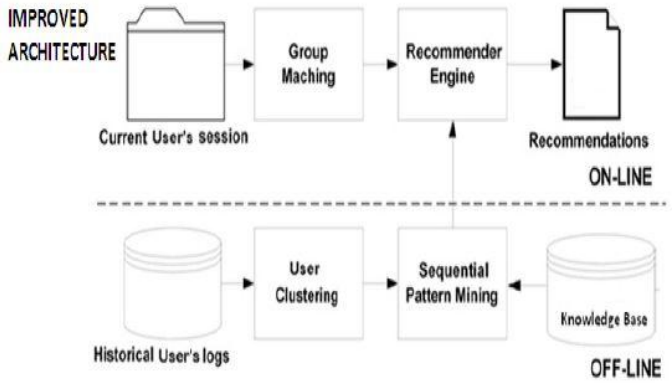

Fig 2. Online & Offline Phase Architecture[3]

Ravi Bhushan and Rajender Nath [4] proposed architecture can be divided into two main phases Back end and Front end. In the back end phase, there are main two modules: Data pre-processing and sequential pattern mining. The block diagram of recommendation system is given below. Back End Phase consists of two modules and modules are data preprocessing and sequential pattern mining. In the front end phase, URL request of the user is processed by search engine and captures the recommended list of web pages relevant to user query and then rank updating algorithm is applied on them.

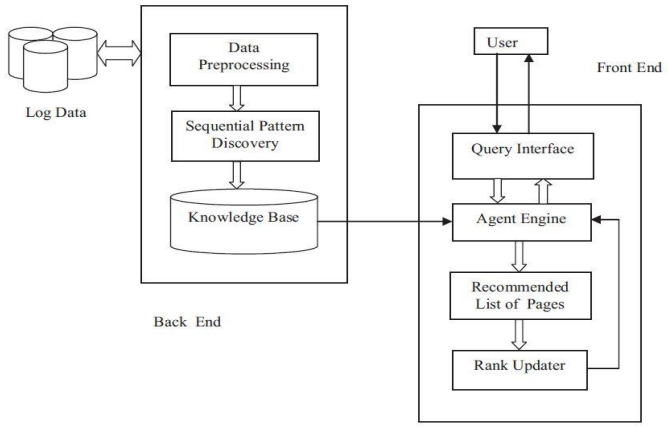

Fig 3. Back End & Front End Architecture[4]

V. Diviya Prabha, R. Rathipriya[5] proposed Gravitational Search Algorithm (GSA) is used to propose a new biclustering algorithm to extract the highly correlated pattern from the optimal bicluster. This GSA is based on the Newtonian gravity: ‘‘Every particle in the universe attracts every other particle with a force that is directly proportional to the product of their masses and inversely proportional to the square of the distance between them”. The clustering techniques work good for small dataset value but work poor for large data sets and if web data is huge that groups similar users under all pages. Besides, algorithm cannot overlap for clusters that are generated i.e. user belong to one cluster may participate in many other clusters with different conditions. To overcome these problems biclustering technique is introduced in the literature. The bicluster are defined to be a set of users and a set of pages where similar users are grouped under specific pages.

IV. CONCLUSION

With the growth of web-based applications, there has been increasing research interest in the discovery and the analysis of web usage patterns. Understanding the browsing behavior of users and applying the discovered knowledge may provide potential increase to the quality of browsing experience. In this work we discuss recent WUM approaches for mining usage patterns. We described the overview of a general WUM process and we highlight recent works done in mining different types of frequent patterns.

REFERENCES

[1] Mr.M.Saravanan, Dr.V.L.Jyothi, “ **A Novel Approach for Sequential Pattern Mining By Using Genetic Algorithm** ” , International Conference on Control, Instrumentation, Communication and Computational Technologies (ICCICCT) ,978 -1-4799-4190-2/14/$31.00 ©2014 IEEE , pg no :284-288

[2] R.Rathipriya , Dr. K.Thangavel , J.Bagyamani, “**Evolutionary Biclustering of Clickstream Data**” , IJCSI International Journal of Computer Science Issues, Vol. 8, Issue 3, No. 1, May 2011 ,ISSN (Online): 1694-0814 , pg no: 341-347

[3] Hiral Y. Modi, Meera Narvekar, “**Enhancement Of Online Web Recommendation System Using A Hybrid Clustering And Pattern Matching Approach**” , 2015 International Conference on Nascent Technologies in the Engineering Field (ICNTE-2015), 978-1-4799-7263-0/15/$31.00 ©2015 IEEE

[4] Ravi Bhushan and Rajender Nath, **“Recommendation of Optimized Web Pages to Users Using Web Log Mining Techniques” ,** 978-1-4673-4529-3/12/$31.00c 2012 IEEE, pg no :1030-1033

[5] V. Diviya Prabha, R. Rathipriya, “**Biclustering of Web Usage Data Using Gravitational Search Algorithm**”, Proceedings of the 2013 International Conference on Pattern Recognition, Informatics and Mobile Engineering, February 21-22, 978-1-4673-5845-3/13/$31.00©2013 IEEE, pg no: 500-505

[6] Rosli Omar, Abu Osman Md Tap, Zainatul Shima Abdullah, “**Web Usage Mining: A Review of Recent Works**”, date : 11/08/2015 time: 10:58 PM

[7] Pablo A. D. de Castro, Fabrício O. de França Hamilton M. Ferreira and Fernando J. Von Zuben, “**Evaluating the Performance of a Biclustering Algorithm Applied to Collaborative Filtering – A Comparative Analysis**”, Seventh International Conference on Hybrid Intelligent Systems, 0-7695-2946-1/07 $25.00 © 2007 IEEE DOI 10.1109/HIS.2007.55, pg no: 65-70

[8] Zhongyun Ying, Zhurong Zhou, Fengjiao Han and Guofeng Zhu, “**Research on Personalized Web Page Recommendation Algorithm Based on User Context and Collaborative Filtering**”, 978-1-4673-5000-6/13/$31.00 ©2013 IEEE, pg no :220-224

[9] Ruimei Lian, “**The Construction of Personalized Web Page Recommendation System in E-commerce**”, 978-1-4244-9763-8/11/$26.00 ©2011 IEEE, pg no :2687-2690

[10] Ming Hu , Guannan Zheng , and Hongmei Wang, "**Improvement and Research on AprioriAll Algorithm of Sequential Patterns Mining**", 2013 6th International Conference on Information Management, Innovation Management and Industrial Engineering, 978-1-4799-0245-3/13/$31.00 ©2013 IEEE, pg no :158-161

[11]Show-Jane Yen, Yue-Shi Lee, Chiu-Kuang Wang, Jung-Wei Wu, "**The Combinations of Frequent Pattern Tree and Candidate Generation for Mining Frequent Patterns**", 2008 Second International Conference on Future Generation Communication and Networking Symposia, 978-0-7695-3546-3/08 $25.00 © 2008 IEEE DOI 10.1109/FGCNS.2008.68, pg no :43-45