Pattern Discovery Techniques in Online Data Mining

Madhur Aggarwal, Anuj Bhatia

***Abstract*— As Web is the largest collection of information; users might devote extra time period on the web outcome the appropriate data or services they are concerned with. The data available is in form of structured (relational) and text data. Therefore, different kinds of data model can be implementable with web data for pattern discovery. Web mining is a data mining tool where the web related data is evaluated for pattern discovery and user navigation pattern. Additionally, according to the nature of data, the kind of mining is also changed. Pattern discovery is used to make a Web site additional responsive to the exclusive and specific desires or requirements of each individual user or set of users. This paper provides a detailed analysis of various approaches of pattern discover in data mining based on different domains with their advantages and limitations. A brief comparison has been made between the different techniques based on certain parameters.**

***Index Terms*— web mining, web usage mining, log analysis, data models, pattern discovery.**

I. INTRODUCTION

The internet is flooded with a lot of useful and useless information. It is very hard to define useful information for a particular user which is varying from time to time. The useful information of one particular time may not be useful on different time or a different situation. The web itself is concerning day by day with newer technologies. Since internet is utilizing free style medium that accepts structured, non-structured, ordered, non-ordered format to provide an information in the web, finding not only the relevant information but to plan them according to the interest of a user is also a key challenge today and is known as web personalization.

"Pattern discovery is the methodology of tailoring a website or content [1] of website to the requirements of every individual user or group of users or organization, taking benefit of the information and web services attained through the exploration of the customer's navigational performance [2]". Pattern Discovery is used to provide services to each specific user in a tailored manner. The propagation of data on the internet has ended the personalization system an obligation. The discovered method must have the capability to resolve the additional data difficulties and let the customers practice at least exertion to find the data they require [3].

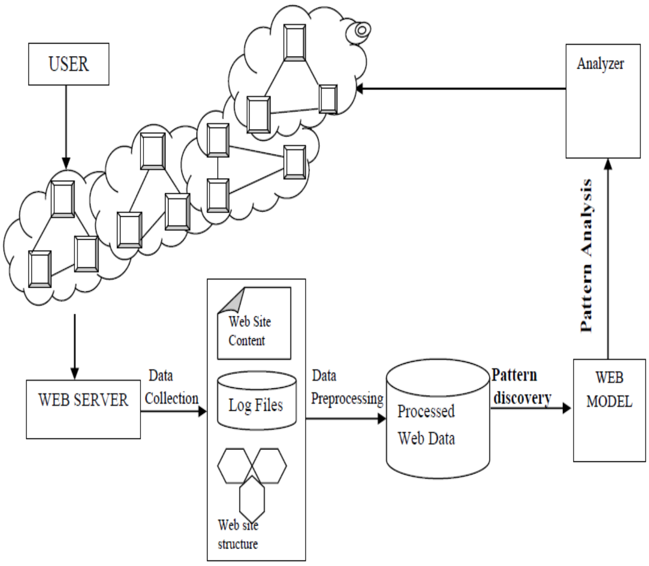

Fig.1 Pattern Discovery Structure

When the customers work on the website their actions can be categorized into two wide sets: browsing and searching. If the customers want to search data on internet, they want to deliver the web scheme with a search demand. If the customers don't provide the web system a precise search request, the system would reoccurrence huge volume of in appropriate data which can't fulfill the requirement of the user. If the customers' request is explicit, searching data on the web might turn out to be simple [4]. To deliver better recommendation to the user the web personalization system's developers must identify what the customers' behavior when they surf on the web.

Therefore, according to data availability on web, web mining can be categorized into three different manners, Web usage mining, Web content mining and Web structure mining.

Web usage mining: the web usage mining allows finding patterns from Web access information. This usage data provides the paths and user access patterns leading to accessed Web pages. This information is often gathered automatically via the Web servers.

Web content mining: the web content mining is also known as text mining. In content mining applications the scanning and mining of text, pictures and graphs of a Web page is performed. That may help to determine the consequence of the content.

Web structure mining: web structure mining is a tool, which is used to recognize the connection between web pages. This organization of data is discoverable by the condition of web structure schema through database techniques for Web pages. This kind of data analysis allows a search engine to pull data concerning to a search query directly to the connecting Web page from the Web sites.

Madhur Aggarwal, B.Tech, IT from Bharati Vidyapeeth';s College of Engg., Associate Technology at Sapient Consulting Pvt Ltd.

Anuj Bhatia, B.Tech. ECE from Graphic Era University, Software Analyst at Accenture Services Pvt Ltd.

II. FORMATS OF DATA

The web access information can be found in different places. Between origin servers to the client end, this access information is organized in different formats that are listed in this section.

Proxy Servers: A proxy server is a software system. That is basically implemented by an organization that is connected to the Internet. Therefore, proxy servers are acts as an intermediary between a host and the Internet connectivity. Using this application the concerning organization can ensure security, caching services and administrative control. Proxy servers can also be a valuable source of usage data. A proxy server also manages access logs, in similar format to Web servers, this access log help to record Web page requests and responses from the web servers.

Client Side Data: Client side data are composed from the host. That is currently accessing the Websites. To collect information directly from the client end, such as the time that the user is accessing and leaving the Web site, a list of sites visited before and after the current site a client agent may helpful.

Client side data are more reliable than server side data. On the other hand, the use of client side data acquisition methods is also challenging. The main problem is that, the different agents accumulating information. That affects the client's system performance. These processes are introducing additional computational and resources overhead when a user tries to access a Web site.

Cookies: In addition to the use the web access log files, a different method often used in the collection of data is the tracking of cookies. Cookies are short strings distributed by the Web server and held by the client's browser for future use. This data is mainly used to track browser visits. By using cookies, Web server can store its own information at the client's machine. Commonly this information is a unique ID that is created by a Web server, by which next time user visits can be realized. Although the maximum size of a cookie cannot be larger than 4 Kbytes therefore it can only store a small amount of information. Additionally, many different cookies may be allocated to a single user. In addition of that, the users may choose to disable the browser option for accepting cookies, Due to privacy and security concerns.

Server Log Files: Server side data are collected at the Web servers of a web site. The web server automatically generates the log file when a user request is made from that server. These logs store Web pages information that is accessed by the visitors of the site. Most of the Web servers support as a default option the Common Log File Format, which includes information about the IP address of the client making the request, the host name and user name, the time stamp of request, file name that is requested, and the file size. The Extended Log Format (W3C), which is supported by Web servers such as Apache and Netscape, and the similar W3SVC format, supported by Microsoft Internet Information Server, include additional information such as the address of the referring URL to this page, i.e., the Web page that carried the visitor to the site, the name and version of the browser used and the operating system of host machine.

III. RELATED WORK

A. Parameters

1. Category Based [8]

In category based, there are two approaches, the first approach is collaborative filtering patterns, and this permits customers to take benefit of other customers' interactive actions based on a degree of likeness between them. Another approach is ruled based pattern; rather than matching customers' response to the web content or summaries of other customers, this model match that query to some fixed rules or conventions, about customer performance.

The following components were considered to implement all the above logic:

- System Logger
- Category Generator
- Customizer

The system logger is intended to gather customers' net usage information. Log files gather customers' visiting counts on every hyperlink on the Web pages. By applying some data mining technique we can excerpt data from the log. Category Generator can categorize the customers into different groups on the basis of log data. Category Generator can identify which customer belongs to which group.

In this approach, author proposed a technique of pattern discovery system stretched from the exploration of classical personalization schemes and associated knowledge. A novel system logger is intended in this approach to store all the content or item openly retrieved by an individual customer, and with the help of category generator, which splits the content into various categories and provide the most appropriate result to the user.

2. Fuzzy Logic Based [9]

In this approach, author proposed a method based on content-based product filtering instead of collaborative filtering and rule based product filtering. Two information segments and two processing segments are considered in this approach. Information module collect the user and service information and processing modules are used to measuring client liking and product filtering, which are the key mechanisms in this personalization method.

The preference learning is supported by the fuzzy logic method which deals with the vague data or facts from the user's activities. The suggested method provides another concept for personalization that adds fuzzy logic for measurement of users' likeness. Fuzzy sets are described as a numerical method to signify and deal with ambiguity or unsure in this area which is depending on membership functions. The membership function describes how each individual fact from input mapped to a membership value in the interval [0, 1]. The proposed method deals with the vagueness of users' activities; the suggested system produced the most appropriate and meaningful value based on the user's behavior and their access time. To produce appropriate membership functions for fuzzy logic is big stimulating issues in fuzzy systems strategy. It is difficult task because it openly affects the correctness of fuzzy logic method.

B. Algorithms

1. Apriori Algorithm [10]

It searches for large itemsets during its initial database pass and uses its result as the seed for discovering other large datasets during subsequent passes. Rules having a support level above the minimum are called large or frequent itemsets and those below are called small itemsets. The algorithm is based on the large itemset property which states: Any subset of a large itemset is large and any subset of frequent item set must be frequent.

The first algorithm for mining all frequent itemsets and strong association rules was the AIS algorithm by [3]. Shortly after that, the algorithm was improved and renamed Apriori. Apriori algorithm is, the most classical and important algorithm for mining frequent itemsets.The Apriori algorithm performs a breadth-first search in the search space by generating candidate k+1-itemsets from frequent k itemsets. The frequency of an item set is computed by counting its occurrence in each transaction.

Apriori is an influential algorithm for mining frequent itemsets for Boolean association rules. Since the Algorithm uses prior knowledge of frequent item set it has been given the name Apriori. It is an iterative level wise search Algorithm, where k itemsets are used to explore (k+1) itemsets. First, the set of frequents 1- itemsets is found. This set is denoted by L1. L1 is used to find L2, the set of frequent 2-itemsets, which is used to find L3 and so on, until no more frequent k-itemsets can be found. The finding of each Lk requires one full scan of database.

There are two steps for understanding that how Lk-1 is used to find Lk:-

The join step:-

To find Lk, a set of candidate k-itemsets is generated by joining Lk-1 with itself. This set of candidates is denoted Ck.

The prune step:-

Ck is a superset of Lk, that is, its members may or may not be frequent, but all of the frequent k-itemsets are included in Ck.

A scan of the database to determine the count of each candidate in Ck would result in the determination of Lk.Ck, however, can be huge, and so this could involve heavy computation to reduce the size of Ck.

2. FP-Tree [12]

A frequent-pattern tree (or FP-tree in short) is a tree structure. It consists of one root labeled as “null”, a set of item-prefix subtrees as the children of the root, and a frequent-item-header table. Each node in the item-prefix subtree consists of three fields: item-name, count, and node-link, where item-name registers which item this node represents, count registers the number of transactions represented by the portion of the path reaching this node, and node-link links to the next node in the FP-tree carrying the same item-name, or null if there is none[2].

Input: A transaction database DB and a minimum support threshold.

Output: FP-tree, the frequent-pattern tree of DB.

FP-growth is a well-known algorithm that uses the FP tree data structure to achieve a condensed representation of the database transactions and employs a divide and-conquer approach to decompose the mining problem into a set of the above problem by Reducing passes, Shrinking number of candidates and facilitating support counting of candidates. An FP-tree-based pattern-fragment growth mining method is developed, which starts from a frequent length-1 pattern (as an initial suffix pattern), examines only its conditional-pattern base (a “subdatabase” which consists of the set of frequent items co-occurring with the suffix pattern), constructs its (conditional) FP-tree, and performs mining recursively with such a tree. The FP-growth algorithm is one of the fastest approaches for frequent item set mining. The FP-growth algorithm uses the FP-tree data structure to achieve a condensed representation of the database transaction and employees a divide-and conquer approach to decompose the mining problem into a set of smaller problems. In essence, it mines all the frequent itemsets by recursively finding all frequent itemsets in the conditional pattern base which is efficiently constructed with the help of a node link structure.

A prefix tree is a data structure that provides a compact representation of transaction data set. Each node of the tree stores an item label and a count, with the count representing the number of transactions, which contain all the items in the path from the root node to the current node. The frequent items are computed as in the Apriori algorithm and represented in a table called header table. Each record in the header table will contain the frequent item and a link to a node in the FP-Tree that has the same item name. Following this link from the header table, one can reach all nodes in the tree having the same item name. Each node in the FP-Tree, other than the root node, will contain the item name, support count, and a pointer to link to a node in the tree that has the same item name.

The main components of FP tree

- It consists of one root labelled as “root”, a set of item prefix sub-trees as the children of the root, and a frequent-item header table.
- Each node in the item prefix sub-tree consists of three fields: item-name, count, and node-link, where item name registers which item this node represents, count registers the number of transactions represented by the portion of the path reaching this node, and node link links to the next node in the FP tree carrying the same item-name, or null if there is none.
- Each entry in the frequent-item header table consists of two fields, (1) item-name and (2) head of node link, which points to the first node in the FP-tree carrying the item-name.

IV. COMPARATIVE ANALYSIS

In this section, we have discussed comparison between above approaches based on different parameters. We have focused on techniques in corresponding approach and advantage, limitation as shown accordingly in table.1.

Algorithm/ Parameter	**Advantage**	**Limitation**
Apriori Algorithm	Searches for large itemset	Full scan require for single itemset
FP-Tree	Scan frequently	More complex for non -frequent item
Category Based	User can take the benefit of other users’ similar interest	Rule based result depend upon developer perception
Fuzzy Logic Based	Deal with uncertainty and ambiguity for better result	Correctness of fuzzy system is difficult

Table.1 Comparative Analysis

V. CONCLUSION

We have discussed algorithm and approaches for pattern discovery based on different domains. They have some strength and weaknesses, but the motive of these work are to make more accurate Web recommendation and provide relevant information and services to each individual user at different point of time by these systems. Some methods are based on content of the web page and users' interest and some of the algorithm includes clustering and data mining techniques. A comparative analysis on the basis of certain parameters a brief comparison is being provided among the all discussed approaches.

REFERENCES

[1] N. Sael, A. Marzak, and H.Behja, "Web Usage Mining Data Preprocessing and Multi Level Analysis on Moodle," IEEE, 2013.

[2] N. Lakshmi, R. S. Rao, and S. S. Reddy, "An Overview of Preprocessing on Web Log Data for Web Usage Analysis,"International Journal of Innovativ Technology and Exploring Engineering (IJITEE), Volume-2, Issue-4, March 2013.

[3] H. peng, "Discovery of Interesting Association Rules on Web Usage Mini ng," International Conference. 2010.

[4] J. Han, J. Pei, and Y. Yin, "Mining frequent patterns without candidate generation: A frequent-pattern tree approach; Data Mining and Knowledge," 2003.

[5] R. Agrawal and R. Srikant, "Fast Algorithms for Mining Association Rules", VLDB. Sep 12-15 1994, Chile, 487-99, pdf, ISBN 1-55860-153-8.

[6] S. Kumar and K.V. Rukmani, "Implementation of Web Usage Mining Using APRIORI and FP Growth Algorithms," 2010.

[7] J. Srivastava, R. Cooley, M. Deshpande, PN. Tan, "Web usage mining: discovery and applications of usage patterns from web data," Vol. 1, No.2, 2000, pp.12–23.

[8] C.C. Lee and W. Xu, "Category-Based Web Personalization System," International Conference on Web Information Systems and Technologies, IEEE 2001, pp.1372-1377.

[9] B.Hua, K. Wai Wong and C.C.Fung, "Fuzzy Logic Based Product Filtering for Web Personalization In E-Commerce," IEEE 2007.

[10] S. Veeramalai, N. Jaisankar, and A.Kannan, "Efficient Web Log Mining Using Enhanced Apriori Algorithm with Hash Tree and Fuzzy" in 2010.

[11] J. Srivastava, R. Cooleyz, M. Deshpande, and P. Tan proposed "Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data," 2000.

[12] J. Han, J. Pei, R. Mao "Mining Frequent Patterns without Candidate Generation" in 2004.

AUTHORS

Madhur Aggarwal, Madhur Aggarwal, B.Tech, IT from Bharati Vidyapeeth's College of Engineering, Associate Technology at Sapient Consulting Pvt Ltd.

Anuj Bhatia, Anuj Bhatia, B.Tech. ECE from Graphic Era University, Software Analyst at Accenture Services Pvt Ltd.