A Survey: Web Mining, Web Log, Pattern Discovery

[1] V.Deepa, [2]S.Ponmalar
[1]Assistant Professor, [2]Assistant Professor,
Department of Computer Science
PSGR Krishnammal College for Women, Coimbatore, India.

Abstract: The term Web mining has been used in two different behaviors. The first, called Web content mining in this paper is the procedure of information discovery from sources across the World Wide Web. The second, called Web usage mining, is the procedure of mining for user browsing and access patterns. They can even study about the visitor's activities through the web analysis and find patterns of the visitor's actions. This kind of web analysis not only involves the modify and interpretation of the web log records to locate the hidden information or predictive pattern by the data mining and knowledge discovery technique, but also offers an immense prospect united with the web mining. Pattern recognition is seen as a foremost challenge within the field of data mining and knowledge discovery. For the work in this paper, we have analyzed mention various stages of web mining, what is web log and a range of widely used algorithms for finding numerous patterns with the purpose of discovering how these algorithms can be used to obtain numerous patterns over large transactional databases.

IndexTerms **- Web Mining, Web Log, Pattern Discovery**

I. INTRODUCTION TO WEBLOG

A weblog, occasionally written as web log or Weblog is a Web site that consists of a series of entries prearranged in reverse chronological order, repeatedly updated on frequently with new information about particular topics. The information can be written by the site owner, gleaned from other Web sites or other sources, or contributed by users.

A weblog often has the excellence of being a kind of "log of our times" from a particular point-of-view. Generally, weblogs are stanch to one or several subjects or themes, usually of newsworthy interest, and in general, can be thought of as rising commentaries, individual or collective on their particular themes. A weblog may consist of the recorded ideas of an individual (a sort of diary) or be a multifaceted collaboration open to anyone.

There are a number of variations on this idea and new variations can easily be make-believe, the meaning of this term is appropriate to gather additional connotations with time. A popular weblog is Slashdot.org, the invention of programmer and graphic artist Rob Malden and several colleagues. Slashdot.org carries conversation threads on many subjects including: Money, Quake (the game), Netscape, Sun Microsystems, Hardware, and Linux. Slashdot.org solicits and a post motivating stories reported by contributors, includes a link to the story, and manages the threads of the subsequent discussion by other users.[9] Another illustrious weblog is Jorn Barger's Robot Wisdom Log, which is more of collected works of daily highlights from other Web sites. Jessamyn West's librarian.net is a daily log of items interesting to librarians and possibly others, besides.

As a layout and content come within reach of for a Web site, the weblog seems popular because the viewer knows that amazing changes every day, there is a personal attitude, and on some sites, there is a prospect to collaborate or react with the Web site and its participants.

Weblog is the forename of a software invention from South Korea that analyzes a Web site's access log and reports the number of visitors, views, hits, most frequently visited pages, and so forth.

II. WEB MINING

Web data mining is an promising research area where mining data is an imperative task and various algorithms has been projected in order to solve the various issues correlated to the web mining in obtainable dataset. This paper focuses the conception of data mining and FP-Growth algorithm. As for FP-Growth algorithm, the efficiency is limited by internal memory size because mining process is on the base of large tree-form data structure. This study work concentrates on web usage mining and in particular focuses on discovering the web usage patterns of web sites from the server log files. This paper finds the system to work with the proposed procedure which can be possible to remove the disadvantage of constraint of the existed technique in the web mining area.

The various web usages mining technique can supplementary work on various scientific area, medical area and social media application to move toward for the research and security related area.

2.1 Stages in Web Mining For Pattern Discovery

Data Preprocessing

The data should be preprocessed to progress the effectiveness and ease of the mining process. The main assignment of data preprocessing is to trim noisy and immaterial data, and to reduce data volume for the pattern discovery segment [8]. Field mining and data cleaning algorithms parse the web log records unscrambling the fields and exclusion.

Pattern discovery

Few techniques to determine patterns from preprocessed data are scheduled like converting IP addresses to domain names, filtering, dynamic site analysis, cookies, path analysis, association rules, chronological patterns, clustering, decision trees etc.

Pattern Analysis

Analysis such as the occurrence of visits per document, most recent visit per document, who is visiting which documents, incidence of use of each hyperlink, and most recent use of each hyperlink. The frequent techniques used for pattern analysis are visualization techniques, OLAP techniques, Data & Knowledge Querying, Usability Analysis.

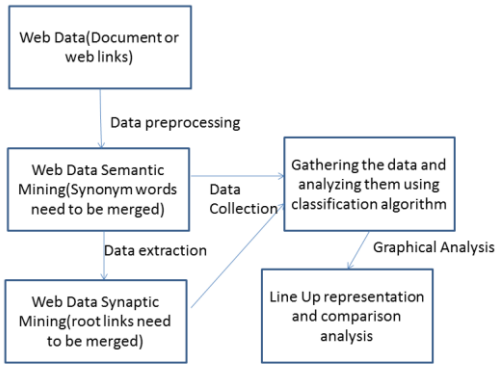


Fig 1: Pattern Analysis Techniques

III. METHODOLOGY OF WORK

Here this paper momentarily describes a technique to discovered frequent item pattern.

A. Semantic Mining

A Web mining from the creep is done first ,technique extracting the information from the web based on the parallel type of object and their ease of use in semantic manner ,the data is been extracted and use to generate Entropy.

B. Synaptic Mining

In this algorithm, the patterns are categorized according to the length executed on network model. Patterns will form a lattice based on the pattern-length and pattern-frequency.

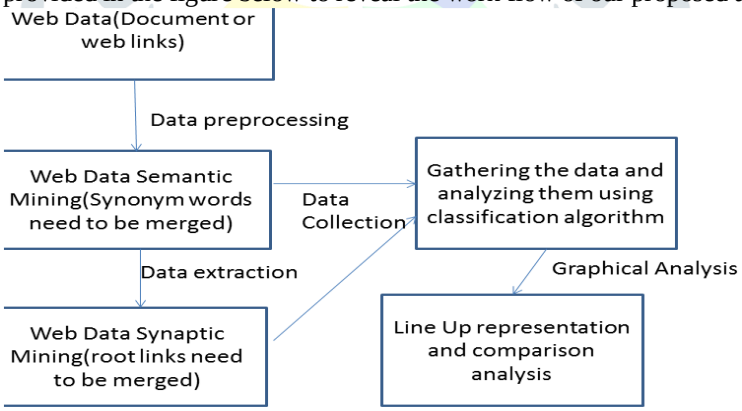
Lattice Construction: The essential constituent of the lattice is an atom i.e. single page. Each atom or page stands for length-1 prefix correspondence class [7]. Commencement from bottom elements the frequency of upper elements with length n can be calculated by using two n-1 length patterns belonging to the same class.

C. Applying *Line Up* on Entropy and Mined Data

The result experimental from the assorted semantic data and user can optimize according to the visualization.

Line-up is a technique which provides a process for the ranking optimization of data which make available the post ranking inference and ranking using different attributes, which offer re-ranking of data using Line up procedure.

Overall process of method is provided in the figure below to reveal the work flow of our proposed architecture.



3.1 Problem Formulation

Today the World Wide Web is admired and interactive medium to allocate information. The web is huge, diverse, dynamic and amorphous nature of web data. Web data investigate encountered lot of challenges for web mining. Information user could encounter following challenges when interact with web.

1. Finding appropriate information- People whichever browse or use the search service when they want to find precise information on the web. Today's search tools have problems like low accuracy which is due to insignificance of many of the search results. This results in a complexity in finding the applicable information. Another problem is low recall which is due to lack of ability to index all the information available on the web.

2. Creating new awareness out of the information obtainable on the web. This problem is fundamentally sub problem of the beyond problem. Above problem is query triggered progression (retrieval oriented) but this problem is data triggered procedure that presumes that previously has group of web data and extract potential useful understanding out of it.

3. Personalization of information- When people relate with the web they differ in the contents and presentations they have a preference.

4. Learning about Consumers or individual users- This crisis is about what the customer do and want. Inside this problem there is sub problem [10].

IV. PATTERN RECOGNITION

Frequent pattern mining can be used in an assortment of actual world applications. It can be used in super markets for selling, product assignment on shelves, for endorsement rules and in text penetrating. It can be used in wireless sensor networks specially in

smart homes with sensors emotionally involved on human body or home procedure objects and other applications that involve monitoring of user location cautiously that are subject to crucial conditions or hazards such as gas leak, fire and explosion .These frequent patterns can be used to observe the activities for dementia patients. It can be seen as a significant move toward with the .1

4.1 Algorithms for Pattern Recognition

4.1.1. Apriori Algorithm

Agrawal and Srikant (1994) firstly projected Apriori algorithm. This algorithm is based on Apriori property which states "each sub *(k-1)*-Item set of recurrent *k-Item set* must be recurrent". Two main processes are executed in apriori algorithm: one is candidate production process, in which the hold count of the equivalent sensor items is calculated by scanning transactional record and second is large item set creation, which is generated by pruning those candidates Item sets which has a bear count less than minimum threshold. These processes are iteratively repeated until candidate Item sets or large Item sets become empty. Original database is scanned first time for the candidate set, consists of one feeler item and there *sustain* has counted, then these 1-*Itemset* candidates are pruned by merely removing those items that has an item count less than user individual threshold (in above case threshold=30%). In second pass database is scanned again to create 2-*Itemset* candidates consist of two items, and then over again pruned to fashioned large 2-*Itemset* using apriori assets. According to apriori property each sub 1-*Itemset* of 2 frequent Item sets must be recurrent [1]. This progression ends as in fourth scan of database 4- *Item set* candidate will be pruned and large item set will be empty.

4.1.2. Association Rule Mining (ARM)

Association rule mining searches for appealing relationships amongst items in a given dataset. The discovery of appealing association relationships among enormous amounts of business transaction records can help in many business decision making processes such as catalog design, cross-marketing and loss-leader analysis. ARM must be emphasized to find out the connection rules that assure the predefined minimum support and confidence from a given database. The support for an itemset is distinct as the ratio of the total number of transactions which contain

The itemset to the total number of contact in the database. The support count for an itemset is the total number of transactions which contain the itemset [2].Support and confidence are two key measures for association rule mining.

Support (A=>B) = P (AUB)

Confidence (A=>B) = P (B/A)

The distinctive example of association rule mining is Market Basket Analysis. This process analyzes customer buying habits by verdict associations between the different items that customers place in their "shopping baskets".extracted from the *SOTrieIT* .RARM also have two limitations. It is difficult to use RARM in interactive mining because if the user support threshold d is changed, the whole process will have to repeat. RARM is also notsuitable for incremental mining, as database size is continuously increasing with addition of new transaction.Whole process needs to repeat again and again.

4.1.3. Equivalence Class Transformation (ECLAT)

ECLAT algorithm uses perpendicular database format whereas in Apriori and RARM horizontal data format (*TranscationId, Items*) has been used, in which operation ids are explicitly listed. While in vertical data format (*Items, TransactionId*) Items with their list of connections are maintained. ECLAT algorithm with set intersection property uses depth-first search algorithm. All numerous *Item sets* can be computed with connection of *TID-* list. In first scan of database a *TID transcationId)* list is maintained for each single item. *k*+1 Item set can be generated from *k* Item set by means of apriori property and depth first search computation. (*k+1)-Itemset* is generated by taking intersection of *TID*-set of frequent *k-Item set* [4].This process continues, until no candidate *Itemset* can be found. One benefit of ECLAT algorithm is that to calculate the support of *k+1* large *Item set* there is no need to scan the database; it is because support count information can be obtained from *k* Item sets. This algorithm avoids the overhead of generating all the subsets of a transaction and checking them against the candidate hash tree during support counting.

4.1.4. Frequent Pattern (FP) Growth Algorithm

It's a two-step approach. In first step a frequent pattern tree is constructed scanning database two times. In first pass of database, data is scanned and support count for each item is calculated, infrequent patterns are deleted from the list and remaining patterns are sorted in descending order. In 2nd pass of database, FP Tree is put together. In 2nd step using FP growth algorithm recurrent patterns are extracted from FP Tree. Conditional FP tree base and Conditional FP tree are based on *node link property* and *prefix path property*. Conditional tree constructed for *I5.* Conditional FP tree is constructed for the frequent items of pattern base FP growth algorithm are good in achieving three important objectives. First is that of which is that database is scanned only two times and computational cost is reduced radically. Second major objective is that no candidate's item set are generated. Third main objective is that it uses divide and conquer move toward which consequently reduces the search space [5]. On the other hands FP growth algorithm has one negative aspect. It is complicated to use in incremental mining, as new transactions are added to the database, FP tree needs to be updated and the whole process needs to repeat.

4.1.5. FP Tree Structure

FP tree is a solid data architecture that retained imperative absolutely vital and quantitative information considering common patterns.

The main attributes of Frequent Pattern tree are:

1. It comprises of one root marked as "root", a set of piece prefix sub-trees as the child of the root, and a frequent-item header Chart.
2. one-by-one node in the piece prefix sub-tree comprises of three areas.

4.2 APPLICATIONS

The general goal of Web Usage Mining is to gather interesting information about users navigation patterns .This information can be exploited later to improve the web site from the users viewpoint. The results produced by the mining of Web logs can be used for various purposes.

A. Personalization of web content: Personalizing the Web Experience for a user is the holy grain of many Web based applications. In this area, recommendation systems are the most common application; their aim is to recommend interesting links to products which could be interesting to users. The Web Watcher, Site Helper, Letizia, and clustering have
All concentrated on providing Web Site Personalization based on usage information.

B. System Improvement: Web usage mining provides the key to understanding Web traffic behavior, which can in turn be used for developing policies for Web caching, network transmission load balancing, or data distribution.

C. Web Site Design: Web usage mining provides detailed feedback on user behavior, providing the Website designer information on which to base redesign decisions.

D. Business Intelligence: Mining business intelligence from Web usage data is dramatically important for ecommerce Web-based companies. Customer Relationship management (CRM) can have an effective advantage from the use of Web.

V. Conclusion and Future Work

With the augmentation of Web based application, particularly electronic commerce, there is noteworthy interest in analyzing Web usage data to enhanced recognize Web usage, and relate the acquaintance to better serve users. This has direct to a number of open issues in Web Usage Mining area. In numerous practical applications, due to the foreword of stricter laws, privacy respect represents big confront. In this survey paper, we momentarily explored various applications of web usage mining optional by authors. We also analyzed some problems and challenges of Web usage mining. Besides we consider that the most appealing research area deals with the combination of semantics within Web site plan so to improve the results of Web Usage Mining applications. Efforts in this direction are likely to be the most fruitful in the creation of much more effective Web Usage Mining and personalization systems that a recons stent mergence and proliferation of Semantic web.

References

[1] Sourav S. Bhowmick Qiankun Zhao, "Association Rule Mining: A Survey," Nanyang Technological University,Singapore, 2003.

[2] Jiawei Han · Hong Cheng · Dong Xin · Xifeng Yan, "Frequent pattern mining: current status and future Directions," *Data Mining Knowl Discov*, vol. 15, no. I, p. 32, 2007.

[3] Iqbal Gondal and Joarder Kamruzzaman Md. Mamunur Rashid, "Mining Associated Sensor Pattern for data stream of wireless networks," in *PM2HW2N '13*, Spain, 2013, p. 8.

[4] M.A. Azam and Loo, J. and Naeem, Usman and Khan, S.K.A. and Lasebae, A. and Gemikonakli Azam, "A Framework to Recognise Daily Life Activities with Wireless Proximity and Object Usage Data," in *3rd IEEE International Symposium on Personal, Indoor and Mobile Radio Communication 2012.*, Sydney, Australia, 2012, p. 6.

[5] Imielienskin T. and Swami A. Agrawal R., "Mining Association Rules Between set of items in largedatabases," in *Management of Data*, 1993, p. 9.

[6] Muhammad Asif, Jamil Ahmed "Analysis Of Effectiveness Of Apriori And Frequent Pattern Tree Algorithm In Software Engineering Data Mining" In Ieee 2015.

[7] Ashika Gupta, Rakhi arora, Ranjana sikarwar"Web Usage Mining Using Apriori Algorithm And Improved Frequent Pattern Tree Algorithm in Association Rule"in IEEE ,2015.

[8] Nandita Agrawal, Anand Jawdekar"User-Based Approach For Finding Various Results In Web Usage Mining"in IEEE 2015.

[9] Hong-Yi Chang 1, Yih-Jou Tzang2,*Zih-Huan Hong1 "A Hybrid Algorithm for Frequent Pattern Mining Using MapReduce Framework"in IEEE 2015.

[10] Avadh Kishor Singh, Ajeet Kumar, Ashish K. Maurya" Association Rule Mining for Web Usage Data to Improve Websites"in IEEE 2014.