Web Mining For Customized Search Using Frequent Pattern Growth Algorithm Techniques

[1] *PriyaSen*
Assistant Professor, Dept. of computer science & Engineering, SVCE College, MP, India

Abstract—*Web mining is a powerful new technology with great potential to help users focus on the most important information in their data warehouses. And their tools predict future trends and behaviors, allowing businesses to proactive and based upon the results it is possible to make decisions too. Web Mining takes the concept of Data Mining into the Internet environment. The proposed system combines the strength of existing works its conceptualization is presented with three phases; namely contents extraction, preprocessing and database of mined data. Based on the user query, the content extraction phase uses the search engine to extract and store raw web pages from the Internet. Clustering or Classification identifies profiles with similar characteristics. Association predicts correlations of items, where one set of transactions implies the existence of another. Sequence patterns involve discovering patterns that indicate usage over a period of time.*

Keywords—*Web content mining, Web structure mining, FP (Frequent Pattern), Classification and Regression Trees (CART).*

I. Introduction

Web Mining Techniques

Web Content Mining is the processing of information, or resource discovery, from multiple sources across the Internet. An agent based approach may be used where a software agent acts autonomously in performing the analysis. A database approach may also be used where an enterprising operational data stores and Internet based information is processed into more structured and high level collection of resources.

Web Structure Mining

Web Structure Mining is interested in analyzing the linkages across the Internet and finding out how they are traversed and from these the website designer has to determine the popularity of web pages.

Benefits of Web Mining

From the collected web log file it is possible to determine the users' interest and also their access pattern. From that information through ontology concept or using association

Rules or clustering concept it is possible to personalize the interested information in their future access.

- Over time enterprises will have built up highly valuable customer information and operational data in their data warehouses. Enterprises now have the capability to collect
- Internet server logs, originally used to monitor server operation but now used to monitor customer behavior. This data is non-volatile, customer or product oriented and primarily used for decision support purposes. The data collected from Internet server logs is the ultimate in customer profiling, recording every activity of the customer based upon the mouse click together with any personal details entered by the customer. Adding these two data sources together gives enterprises the ultimate customer information, feedback and profiling systems.
- With this amount of data available from the web log file enterprises they are able to perform direct marketing with the ability to identify and target users with personalized offers and promotions. Web mining allows enterprises to provide enhanced service offerings including personalization, collaborative filtering, and enhanced customer support and product and service strategy definitions.
- By using Web Mining techniques enterprises are able to understand their customers, their buying patterns and behaviors which is essential if they are to retain them as customers. Furthermore, Web Mining provides enterprises with the opportunity to optimize their sites for maximum commercial impact and the ability to provide personalized online content of their web site.
- Web Mining techniques provide the opportunity for developing customer insight at an unprecedented level. The enterprises that make the best use of traditional data sources and Internet sourced data through Web Mining technique will be placed to deliver and succeed with their e-business strategy.

II.Literature Survey

Kumar, T. Vijaya, et al. [4] proposed a model which incorporates website knowledge in web usage mining techniques. In this paper, the authors have introduced a new idea of incorporating available website knowledge for better session construction which would eventually lead to better patterns during pattern discovery. Sharma et al. [9] explains that when applied to Web Usage Mining, association rule mines are used to find associations among web pages that frequently appear together in users' sessions. The authors have proposed different learning web usage patterns and evaluated some interesting measures to evaluate the association rules mined from web usage data. Munibalaji et al. [7] has given a survey of page ranking algorithms and description about Weighted Page Content Rank (WPCR) based on web content mining and structure mining that shows the relevancy of the pages to a given query is better determined, as compared to the Page Rank and Weighted Page Rank algorithms. Rao, T. K., et al. [8] has explained how cloud computing can be used effectively for Web mining in e-commerce

organizations to reduce costs and increase the profitability of the e-commerce enterprise.

III. Pre-Processing Concept

Using the techniques used in Data Mining, Web Mining applies the techniques to the Internet by analyzing server logs and other personalized data collected from customers to provide meaningful information and knowledge. Here we fully concentrate on the pre-processing and analysis of the web log data. This is possible to find out the information about a web site, top errors, valuable visitors of that web site and based upon this information the Web site designers has to improve their sites and correct the errors and store the usable web log database for further mining process. We studied the pre-processing of the web log data, how it was handled in the existing systems and what are the problems faced during pre-processing the web log data and also the existing problems in the pre-processing concept. The researcher collected all these information from the following papers and it was very useful for my research. Feng Zhang and Hui-You Chang [13] analyzed and found immediate solution for the three tasks and there is no period after the "et" in the Latin abbreviation "et al."

Key issues of the Web Usage mining. In general, web Mining can be classified into three domains: Web structure mining, Web content mining and Web usage mining. The three main tasks of Web usage Mining are pre- processing, knowledge discovery and pattern analysis. Though Web Usage Mining is still ranged in the application of traditional data mining techniques, in view of changes in application environment and operated data concerned, some new difficulties have cropped up. Here we make efforts to address such challenges in the three phases and proposed some solutions simultaneously. Here we have taken great efforts to address the key issues in Web Usage Mining and have introduced some solutions. Actually Web Usage Mining, as well as Web Mining, is a new research field, which is still in the development phase and has a long way to go.

Ranieri Baraglia and Paolo Palmerini proposed a WUM system called SUGGEST which analyzed the past user behavior to find out the future access of that user. During their navigation web users leave many records of their activity. This huge amount of data can be a useful source of knowledge. Sophisticated mining processes are needed for this knowledge to be extracted, understood and used. In this paper we propose a Web Usage Mining (WUM) system, called SUGGEST, a designed to efficiently integrate the WUM process with the ordinary web server functionalities. It can provide useful information to make easier the web user navigation and to optimize the web server performance. Here we studied the problem of the realization of a Web Usage Mining system. We proposed behavior and also it is possible to find out the future access of the user.

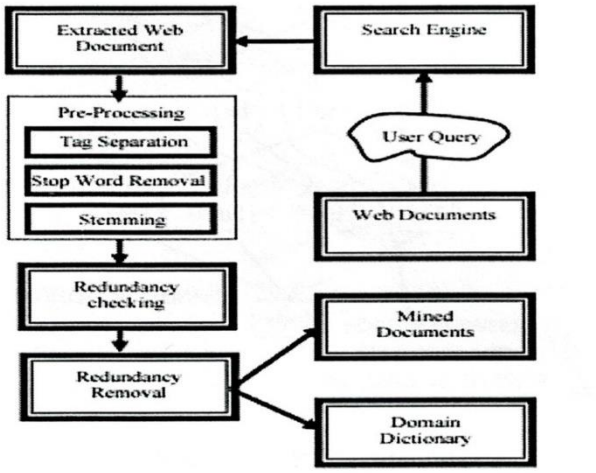
IV.Need and Overview ofthe Pre-Processing Concept

A web log file is a collection of information about the users' accessed web pages in the website. It was not possible to use the raw web log file for mining process because during that process it may lead to wrong (or) inconsistent error and sometimes it may lead to failure.

Also. To overcome that problem, we designed a model for pre-processing the web log for further mining process. Generally web logs may be collected in three sources. They are

- Client side
- Server side and
- Proxy severs

There was some problem in collecting the web log data from client side and proxy side and also these kinds of data were not complete and consistent that was also a main reason of our model to collect the web logs from server side. In general most of the pre-processing algorithm was used for the server side web log data. Here we collected web log from a reputed website and the structure diagram of the website is shown in the following Figure-3- 1. In figure the user start from the home page at the left end of the diagram and moves towards the right using the links to reach the desired page which were needed by the user. In the figure there is a link to access either DDE website or the link to access the Library website. Here the model collected web logs from the three websites (Home website, DDE website, Library website) and each access of the user was recorded in the corresponding web log file separately.



PROPOSED SYSTEM ARCHITECTURE

During the pre-processing time the first step was the Data merging / Data Integration which combines data from multiple sources into a single data store. Here we collected the web logs from the three websites and using the data integration concept they were merged into one and from that the model does the further pre-processing steps.

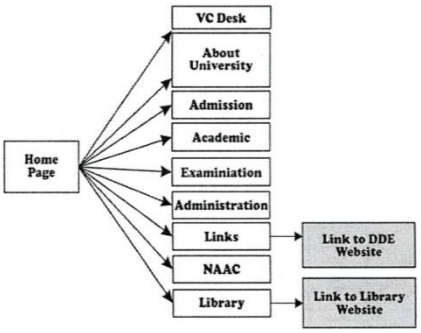


Figure.1 Structure diagram of the reputed website

V. FREQUENT PATTERN MINING WITH THE TRADITIONAL SUPPORT FRAMEWORK

The support framework is designed to determine patterns for which the raw frequency is greater than a minimum threshold. Although this is a simplistic way of defining frequent patterns, this model has an algorithmically convenient property, which is referred to as the level-wise property. The level-wise property of frequent pattern mining is algorithmically crucial because it enables the design of a bottom-up approach to exploring the space of frequent patterns Since the problem of frequent pattern mining was first proposed, numerous algorithms have been proposed in order to make the solutions to the problem more efficient.

FIMI was devoted to implementations of frequent pattern mining for a few years. This site is now organized as a repository, where many efficient implementations of frequent

Pattern mining are available. The techniques for frequent pattern mining started with Priory-like join-based methods. In these algorithms, candidate item sets are generated in increasing order of item set size. The generation in increasing order of item set size is referred to as level-wise exploration. These item sets are then tested against the underlying transaction database and constraint are retained for further exploration. Eventually, it was realized that these prior-like methods could be more systematically the frequent ones satisfying the minimum support explored as enumeration trees. This structure will be

Or other hybrid strategies [13]. One property of the breadth-first strategy is that level-wise pruning can be used, which is not possible with other Strategies. Nevertheless, strategies such as depth-first search have other advantages, especially for maximal pattern mining.

This observation for the case of maximal pattern mining was first stated in [12]. This is because long patterns are discovered early, and they can be used for downward closure-based pruning of large parts of the enumeration tree that are already known to be frequent. It should be pointed out, that for the case where all frequent patterns are mined, the order of exploration of an enumeration tree does not affect the number of candidates that are explored because the size of the enumeration tree is fixed.

Vi. Result Discussion and Performance Study

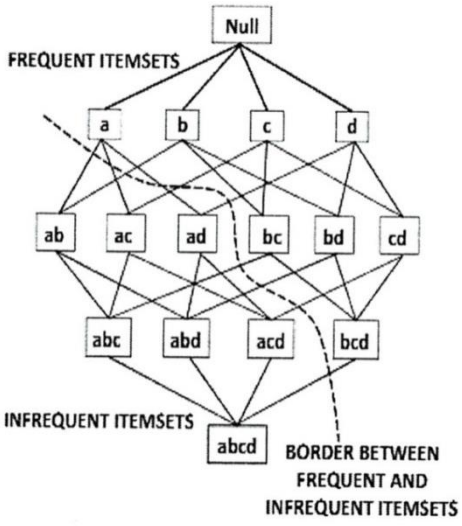


Fig 2 Item sets in frequent mining

The phase of the bucketing operation requires |E (P)| iterations, and each iteration requires 2” operations. Therefore, the total time required by the method is proportional to |E (P)|. When |E(P)| is sufficiently small, the time required by the second phase of post processing is small compared to the first phase, whereas the first phase is essentially proportional to reading the database for the current projection.

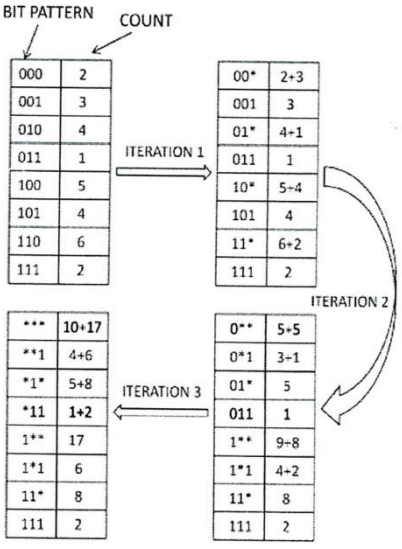


Fig 3 Frequent mining bucketing operation

We have illustrated this phase of bucketing by an example in which |E (P)| = 3. The process illustrated in Fig. illustrates how the second phase of bucketing is efficiently performed.

The exact strings and the corresponding counts in each of the |E (P)| = 3 iterations are illustrated. In the first iteration, all those bits with 0 in

the lowest order position have their counts added with the count of the bit string with a 1 in that position. 2"+" pairwise addition operations take place during this step. The same process is repeated two more times with the second and third order bits. At the end of three passes, each bucket contains the support count for the appropriate item set, where the '0' for the item set is replaced by a "don't care" which is represented by a ''. Note that the number of transactions in this example is 27. This is represented by the entry for the bucket **. Only two transactions contain all three.

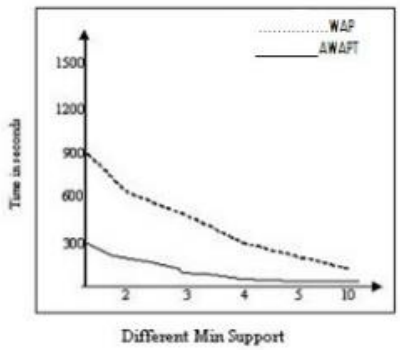
Algorithm 1 (WAP-tree Construction for Web access sequences)
Input: Access sequence database D (i), min support MS (0< MS ≤ 1)
Output: frequent sequential patterns in D (i). Variables: Cn stores total number of events in suffix trees, a stores whether a node is ancestor in queue.
Begin
1. Create a root node for T;
2. For each access sequence S in the access sequence database AWAPT do
a) Extract frequent subsequence S1 =S1 S2 …...Sn, WHERE S1 (1<=I<=n) are events in S1 .Let current node point to the root of T.
b) for i=1 to n do , if cuurent_node has a child labeled Si by 1 and make current node point Si , else create a new child node(S1:1),make current node point to the new node, and insert it into the Si queue 3. Return (T);

EXPERIMENTAL EVALUATION AND PERFORMANCE STUDY.

In this section, we report our experimental results on the performance of AWAPT in comparison with WAP Tree and FS-Tree. It shows that AWAPT outperforms other previously proposed methods and is efficient and scalable for mining sequential patterns in large databases.

Algorithms	Time in sec's at different supports				
	2	3	4	5	10
WAP	750	510	330	280	150
AWAPT	230	160	110	95	48

Execution times for dataset at different Minimum supports.



Execution times trend with different minimum supports.

In summary our performance study shows that AWAPT is more efficient and scalable than WAP Tree and FS-Tree, Whereas WAP tree is faster than FS -tree when the support threshold is low, and there are many long patterns. The AWAPT algorithm eliminates the need to store numerous intermediate WAP trees during mining. Since only the original tree is stored, it drastically cuts off huge memory access costs, which may include disk I/O cost in a virtual memory environment, especially when mining very long sequences with millions of records. This algorithm also eliminates the need to store and scan intermediate conditional pattern bases for reconstructing intermediate WAP trees.

V.Conclusions

We have developed a novel, scalable, and efficient frequent sequential pattern mining method, called AWAPT. Our systematic performance study shows that AWAPT mines the complete set of patterns and is efficient and runs considerably faster than both based WAP Tree and FS-Tree algorithms. With standard tools for web content mining, there is opportunity for extracting only the relevant text from web while unrelated textual noise like advertisements, navigational elements, contact and copyright notes are reliably suppressed. The reported research hybridized graph theoretic and genetic algorithm to formulate a web content mining technique for achieving this purpose. The new technique provides timely search and discovery from large web datasets and experimental results had shown its superiority over other techniques. These suggest the new technique will be very useful in areas where knowledge discovery, web structure and web analytics are required. It is of note that the applicability of the new technique on complex and large number of parameters has not been investigated. Limitation of our model was, here the model the user's uniform interest pattern based upon the tree traversal results.

REFERENCE

[1] Liu B., Structured data extraction: Wrapper generation and Web Data Mining, Editorial Issues on Web content Mining. SIGKDD Explorations —Vol. 6, 2005, pp. 363 — 423

[2] Abdelhakim H., Khentout C. and Djoudi M., Overview of Web Content Mining Tools, The International Journal of Engineering and Science (IJES), Vol. 2, 2013.

[3] Marghny M. H. and Ali A. F., Web mining based on genetic algorithm, Proceedings of AIML '05 Conference, Cairo, Egypt, 2005, pp 19-21

[4] Kumar, T. Vijaya, et al. "A New Web Usage Mining approach for Website recommendations using Concept hierarchy and Website Graph." International Journal of Computer and Electrical Engineering 6.1 (2014)

[5] Zaiane R. O., Introduction to Data Mining: Principles of knowledge discovery in databases, 1999.

[6] Han J. and Kamber M., Data mining: concepts and techniques, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2000

[7] Munibalaji, T., and C. Balamurugan. "Analysis of link algorithms for web mining. “International Journal of engineering and Innovative Technology 1.2 (2012).

[8] Rao, T. K., et al. "Mining the E-commerce cloud: A survey on emerging relationship between web mining, Ecommerce and cloud computing." Computational Intelligence and Computing Research (ICCIC), 2013 IEEE International Conference on.IEEE, 2013.

[9] Eliot.T.S.1972“The Sacred Wood: Essays on poetry and Criticism“, Methuen and Co. Ltd. London

[10] Chen M. S., Han J. and Yu P. S., Data mining: An overview from a database perspective, IEEE Trans. Knowledge and Data Engineering, Vol. 8, 1996, pp. 866- 883.

[11] Mrs. Kirti Tandele, “Web Usage Mining with Improved Frequent Pattern Tree Algorithms”, International Journal of Computer Science and Information Technology
Research, Vol. 3, Issue 2, pp.: (952-958) 2015.

[12] Imielinski T. and Mannila H., A database perspective on knowledge discovery, Communications of ACM, Vol. 39, pp. 58-64.

[13] Cooley R., Mobasher B. and Srivastava J... Web mining: information and pattern discovery on the World Wide Web. Proceedings of 9th IEEE International Conference, pp. 558 — 567, 1997

[14] Ashika Gupta et al. “Web Usage mining using Improved Frequent Pattern Tree Algorithm”, International Conference on Issues and Challenges in Intelligent Computing Techniques (2014)

I.