Handling User Navigation from Weblogs using P-Prefix Span Algorithm

Ms. T. Roobini, Mr. P.G. Om Prakash

Abstract— **Now a Days, People are increasingly using World Wide Web to extract the valuable information retrievals and knowledge discoveries. We know that the amount of online data transaction has been increased. Web Mining Technologies are the right solution for knowledge discovery on the Web. The two most important tasks in information extraction from the web are webpage structure understanding and natural language sentence processing. P-Prefix Span technique is to improve in different aspects like Performance for handling user navigation. In P-Prefix Span algorithm it has sequential pattern extract frequently occurring Inter-Session pattern such that presence of set of items followed by another item in time order. The task of Web Mining is to discover and extract interesting knowledge/pattern from web is classified into three types as web structure mining that focuses on hyperlink structure, web content mining page contents as well as web usage mining that focuses on web logs. Web mining is the base for navigation pattern mining and approach of clustering is used to perform that mining, Usage Mining deals with the discovery and analysis of usage patterns from web data, specifically web logs, in order to improve Web based applications and also to improve the customer relationship management. The algorithm also provide Dependency Modelling determine, if there is any significant Dependencies among Variables in Web.**

Index Terms— **Web Traversal Pattern(WTP), Profile Aggregations based on Clustering Transactions(PACT), Knowledge Discovery in Databases(KDD), Web Intra-Page Informative Structure Mining Based on the Document Object Model(WISDOM).**

I. INTRODUCTION

The World Wide Web contains huge amounts of data. However, we cannot benefit very much from the large amount of raw web pages unless the information within them is extracted accurately and organized well. Therefore, information extraction (IE) plays an important role in web knowledge discovery and management. Among various information extraction tasks, extracting structured Web information about real-world entities (such as people, organizations, locations, publications, products) has received much attention of delay. However, little work has been done towards an integrated statistical model for understanding webpage structures and processing natural language sentences within the HTML elements of the webpage.

Recently, millions of electronic data are included on hundreds of millions data that are previously on-line today. With this significant increase of existing data on the Internet and because of its fast and disordered growth, the World Wide Web has evolved into a network of data with no proper organizational structure. Guessing the users' interests for improving the usability of web or so called personalization has turn out to be very essential and difficult in this situation. Generally, three kinds of information have to be handled in a web site: content, structure and log data. The usage of the data mining process to these dissimilar data sets is based on the three different research directions in the area of web mining: web content mining, web structure mining and web usage mining.

Web search logs include searching activities of users in search engines. Previous studies have shown that search logs can be used in various applications including user satisfaction analysis, page utility estimation, user search interest prediction, query suggestion, web page re-ranking, web site recommendation, etc. Most of previous work analyzed web search logs at session or query level, where a session is defined as "a series of queries by a single user made within a small range of time". However, few of them have considered search logs at task (atomic user information need) level. Web log mining consists of three main steps:

A. Data pre-processing

Web log data pre-processing is a complex process and takes 80% of total mining process. The aim of data pre-processing is to select essential features clean data by removing irrelevant records and finally transform raw data into sessions.

B. Pattern extraction

Pattern mining refers to find the various hidden, potentially useful information from a large amount of pre-processed data. There are different types of data mining approaches are used to extract the patterns such as Frequent Item Discovery, Frequent Sequence Discovery, Frequent Sub-tree Discovery.

C. Analysis of extracted patterns

Analysis and evaluation module is to analyse the credibility and effectiveness of the extracted patterns i.e. how these patterns can be used to analysis and predict a user behaviour, satisfaction, web site ranking and top level related query suggestion.

II. RELATED WORK

A. MinHash

Manuscript received April 25, 2015

Ms. T. Roobini, PG Scholar, Department of CSE, St.Joseph Engineering College, Chennai

Mr. P.G. Om Prakash, Assistant Professor, Department of CSE, St.Joseph Engineering College, Chennai

Extracting structured information from unstructured and/or semi-structured machine-readable documents automatically plays a major role now a days, So most websites are using common templates with contents to populate the information to achieve good publishing productivity, Where WWW is the major resource for extracting the information. In recent days Template detection technique received lot of concentration to improve in different aspects like performance of search engine , clustering and classification of web documents , as templates degrade the performance and accuracy of web application for a machines because of irrelevant template terms.

In this research, we propose to represent a web document and a template as a set of paths in a DOM tree. As validated by the most popular XML query language XPATH, paths are sufficient to express tree structures and useful to be queried. Our template detection method is based on the repetition of text segments which are text nodes in DOM trees of web pages. We use a data structure called the text segment table to maintain the repetition information, i.e. the contents and DFs of text segmentation.

We employed the MDL principle to manage the unknown number of clusters and to select good partitioning from all possible partitions of documents, and then, introduced our extended **MinHash** technique to speed up the clustering process. Experimental results with real life data sets confirmed the effectiveness of our algorithms.

B. User Profiling

Web personalization is the process of customizing a Web site to the needs of specific users, taking advantage of the knowledge acquired from the analysis of the user's navigational behaviour (usage data) in correlation with other information collected in the Web context, namely, structure, content and user profile data. Due to the explosive growth of the Web, the domain of Web personalization has gained great momentum both in the research and commercial areas.

Using such information, they can optimize their site in order to increase sales and ensure customer retention. Apart from Web usage mining, User Profiling techniques are also employed in order to form a complete customer profile. Lately, there has been an effort to incorporate Web content in the recommendation process, in order to enhance the effectiveness of personalization.

C. Profile Aggregations based on Clustering Transactions

Web usage mining, possibly used in conjunction with standard approaches to personalization such as collaborative filtering, can help address some of the shortcomings of these techniques, including reliance on subjective user ratings, lack of scalability, and poor performance in the face of high-dimensional and sparse data. However, the discovery of patterns from usage data by itself is not sufficient for performing the personalization tasks. The critical step is the effective derivation of good quality and useful (i.e., actionable) "aggregate usage profiles" from these patterns.

Our evaluation results suggest that each of these techniques exhibits characteristics that make it a suitable enabling mechanism for different types of Web personalization tasks. The first technique, called PACT (Profile Aggregations based on Clustering Transactions), is based on the derivation of overlapping profiles from user transactions clusters. This later observation also indicates another advantage of usage-based Web personalization over traditional collaborative filtering techniques which must rely on deeper knowledge of users or on subjective input from users (such as book or music ratings).

III. CONCEPT

Design is a meaningful engineering of something that is to be built. Software Design sites at the technical kernel of software engineering. Software design is a process through which the requirements are translated in to a representation of the software i.e. the blue print for constructing software. Design provides us with representation of software that can be assesses for quality. Design is the only way that we can accurately translating a customer's requirements in to a finished software product.

A. Information Pre-processing

In this module, initially prepare the weblog dataset. The dataset contains history of the user browsing details and web search logs contain user behaviours on search engines, like clicks and queries. The prepared raw weblog dataset contains null values and unwanted datasets. Pre-processing technique is used to remove the null values in web search log dataset. After the pre-processing process completed store the data into database.

The first issue in the pre-processing phase in data preparation. Depending on the application, web log data may need to be cleaned from entries involving pages that returned an error or graphics file accesses. In some cases such information might be useful, but in other such data should be eliminated from log file. Furthermore ,crawler activity can be filtered out, because such entries do not provide useful information about the site's usability. Another problem to be met has to do with caching. Accesses to cached pages are not recorded in the web log, therefore such information is missed. Caching is heavily dependent on the client-side technologies used and therefore cannot be deal with easily. In such cases, cached pages can usually be inferred using the referring information from the logs. Moreover, a useful aspects is to perform page view identification, determining which page file accesses contribute to a single page view. Again such a decision is application-oriented.

B. Knowledge Discovery

The project belongs to **"KNOWLEDGE & DATA MINING"** Approach Data mining (the analysis step of the "Knowledge Discovery in Databases" process, or KDD), a relatively young and interdisciplinary field of computer science, is the process that results in the discovery of new patterns in large data sets. It utilizes methods at the intersection of artificial intelligence, machine learning, statistics, and database systems. The overall goal of the data mining process is to extract knowledge from an existing data set and transform it into a human-understandable structure for further use. Besides the raw analysis step, it involves database and data management aspects, data pre-processing, model and inference considerations, interestingness metrics, complexity considerations, post-processing of found structures, visualization, and online updating.

The term is a buzzword, and is frequently misused to mean any form of large-scale data or information processing

(collection, extraction, warehousing, analysis, and statistics) but is also generalized to any kind of computer decision support system, including artificial intelligence, machine learning, and business intelligence. In the proper use of the word, the key term is discovery, commonly defined as "detecting something new". Even the popular book "Data mining: Practical machine learning tools and techniques with Java" (which covers mostly machine learning material) was originally to be named just "Practical machine learning", and the term "data mining" was only added for marketing reasons. Often the more general terms "(large scale) data analysis", or "analytics" - or when referring to actual methods, artificial intelligence and machine learning - are more appropriate.

The actual data mining task is the automatic or semi-automatic analysis of large quantities of data to extract previously unknown interesting patterns such as groups of data records (cluster analysis), unusual records (anomaly detection) and dependencies (association rule mining). This usually involves using database techniques such as spatial indexes. These patterns can then be seen as a kind of summary of the input data, and may be used in further analysis or, for example, in machine learning and predictive analytics. For example, the data mining step might identify multiple groups in the data, which can then be used to obtain more accurate prediction results by a decision support system. Neither the data collection, data preparation, nor result interpretation and reporting are part of the data mining step, but do belong to the overall KDD process as additional steps.

The related terms data dredging, data fishing, and data snooping refer to the use of data mining methods to sample parts of a larger population data set that are (or may be) too small for reliable statistical inferences to be made about the validity of any patterns discovered. These methods can, however, be used in creating new hypotheses to test against the larger data populations.

C. Classification

Classification consists of predicting a certain outcome based on a given input. In order to predict the outcome, the algorithm processes a training set containing a set of attributes and the respective outcome, usually called goal or prediction attribute. The algorithm tries to discover relationships between the attributes that would make it possible to predict the outcome. Classification consists of assigning a class label to a set of unclassified cases.

D. Clustering

Clustering is a useful and important unsupervised learning technique widely studied in literature. The general goal of clustering is to group similar objects into one cluster while partitioning dissimilar objects into different cluster .our clustering framework is to partition an attribute graph G based on both structure and attribute similarities through a unified neighbourhood random walk model on the attribute-augmented graph Ga of G. the objective of clustering is to maximize intra-cluster similarity and minimize inter-cluster similarity.

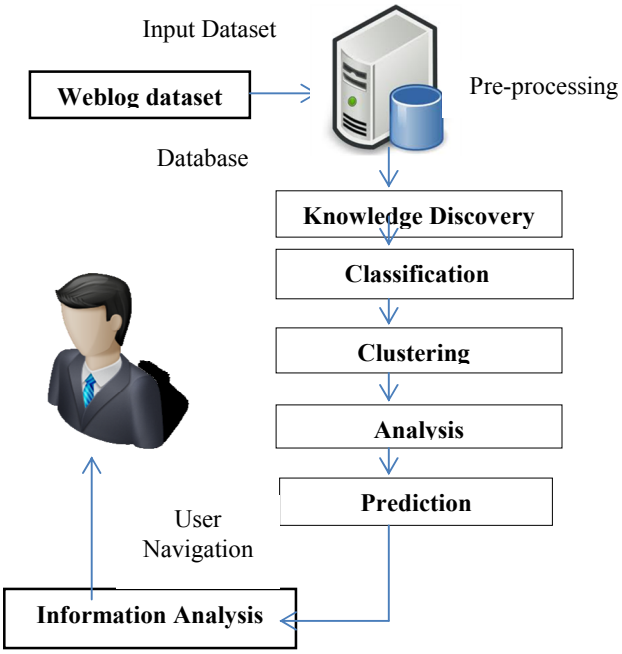


Fig.1. System architecture

E. Information Analysis

In this module, Based on the ranking process classify the weblog dataset into frequent and infrequent item set. Here, frequent item set indicates the user navigation. Then, forward the frequent item set and navigation to the respected user. This the procedure where information stored in web server logs is processed by applying data mining techniques in order to ,Extract statistical information and discover interesting usage patterns, Cluster the users into groups according to their navigation behaviour. Discover potential correlation between web pages and user groups.

This process of extracting information concerning the browsing behaviour of the users can be regarded as part of the user profiling process. It is therefore evident that the user profiling and web usage mining modules overlap. In many cases information provided by a web site is not physically stored in the web site's server. In the case of a web portal , users are interested in information from various web sources. It remains to the web site editors to search the web for content of interest that should consequently be classified into thematic categories. Searching and relevance ranking techniques must be employed both in the process of acquisition of relevant information and in the publishing of the appropriate data to each group of users.

IV. PROPOSED ALGORITHM

P-Prefix span algorithm is used to handle the user navigation from web logs. Dependency Modelling determines if there are any significant dependencies among the variables in the Web. The proposed algorithm can discover frequent sequential patterns with probability of inter arrival time of consecutive items. With the estimated probability of inter arrival time, the algorithm can be employed to re-evaluate the support of searched frequent patterns. Sequential Patterns extract frequently occurring inter-session patterns such that the presence of a set of items s followed by another item in time order. Classification is the technique to map a data item into one of several predefined classes

ALGORITHM APPROACH

The proposed system works based on the **P-Prefix Span algorithm**

Step-1 Start

Step-2 Set *ξ, v, td*

Step-3 <α> = null

Step-4 Scan *S*|<*α*> to generate frequent items

Step-5 Estimate the inter-arrival time probability for each frequent item in a time period *td*

Step-6 Append the frequent item with inter-arrival time probability greater than or equal to *v to* <α> to form a new reliable pattern <α'>

Step-7 Construct *S*|<*α* '>

Step-8 Call P-Prefix Span (<*α* '>,|<*α* '>|, *S*|<*α* '>))

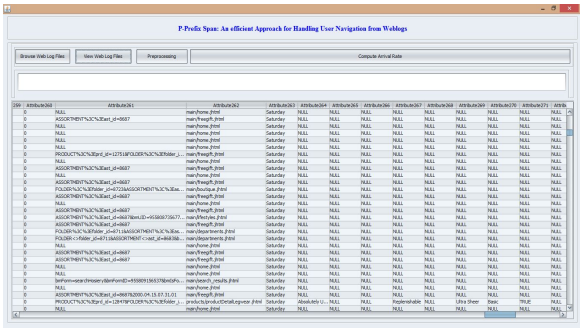
V. DISCUSSION

The main objective of this project is to handle the user navigation from weblogs by using P-Prefix Span Algorithm to provide efficient user handling navigation . The Project is to Analyse the User Behaviour Pattern from Weblogs and reduce the Mining time. This Research work deals with prediction of user web browsing behavior using web mining technique. In the web usage mining, the usage data are analysed and correspond to user navigation throughout the web rather than particular website.

VI. RESULTS

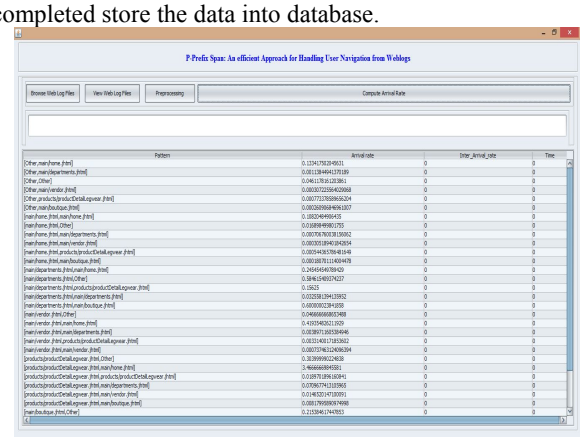
Loading Raw Data

The dataset contains history of the user browsing details and web search logs contain user behaviours on search engines, like clicks and queries. The prepared raw weblog dataset contains null values and unwanted datasets.

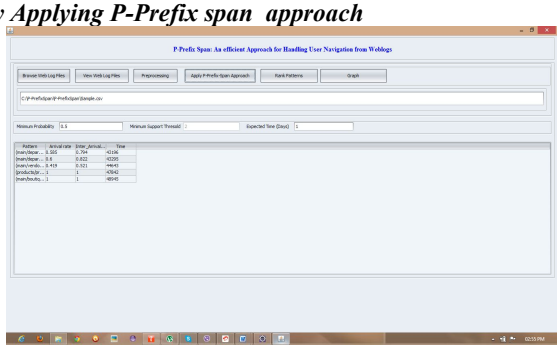


Pre-processed Data

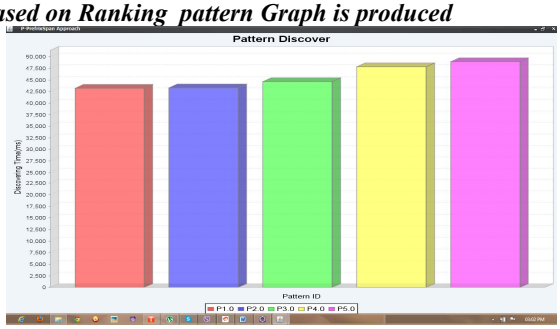
Pre-processing technique is used to remove the null values in web search log dataset. After the pre-processing process completed store the data into database.



By Applying P-Prefix span approach



Based on Ranking pattern Graph is produced



VII. CONCLUSION

This project proposes a new algorithm, P-Prefix Span algorithm for mining frequent patterns, which concentrate on symbolic patterns and numerical analysis simultaneously. The frequent pattern can yield information not only of the order of frequent items but also of the time probability of arrival items. The information, which provides a more detailed description about the behavior of derived patterns, is crucial to decision-makers. According to our proposed algorithm, the arrival rate will be computed based on total number of sub sequence present in projected database. Calculate the inter-arrival rate of selected item using arrival rate and expected, if inter-arrival rate is greater than the user specified minimum probability threshold append the frequent item with selected item, Construct the projected database with append item and selected item, continue the process until specified minimum support value. Based on the arrival rate and inter-arrival rate calculation, rank the weblog dataset into ascending order. Here, frequent item set indicates the user navigation. Then, forward the frequent item set and navigation to the respected user.

VIII. REFERENCES

[1] Aggarwal C.C and H.Wang , Managing and Mining Graph Data. New York,NY,USA: Springer,2010.

[2] Angelova .R and S.Siersdorfer,@A neighbourhood-based approach for clustering of linked document collections,@ in Proc. SDM Conf.,2006,pp.778-779.

[3] Eirinaki .M and M. Vazirgiannis, "Web Mining for Web Personalization," ACM Trans. Internet Technology, vol. 3, no. 1, pp. 1-27, 2003.

[4] Galletta D.F, R. Henry, S. McCoy, and P. Polak, "When the Wait Isn"t So Bad: The Interacting Effects of Website Delay, Familiarity, and Breadth," Information Systems Research, vol. 17,no.1,pp.20- 37, 2006.

[5] Guha S,R.Rastogi, and K.Shim,@CURE: An efficient clustering algorithm for large databases,@ in Proc. ACM SIGMOD Conf., New York, NY, USA,1998,pp.73-84.
[6] Kao.H, J. Ho, and M. Chen, "WISDOM: Web Intrapage Informative Structure Mining Based on Document Object Model," IEEE Trans. Knowledge and Data Eng., vol. 17, no. 5, pp. 614- 627, May 2005.
[7] Kim.C and K. Shim, "TEXT: Automatic Template Extraction from Heterogeneous Web Pages," IEEE Trans. Knowledge and Data Eng., vol. 23, no. 4, pp. 612-626, Apr. 2011.
[8] Mobasher.M, H. Dai, T. Luo, and M. Nakagawa, "Discovery and Evaluation of Aggregate Usage Profiles for Web Personalization, " Data Mining and Knowledge Discovery, vol. 6, no. 1, pp. 61-82, 2002.
[9] Sebastiani.F,"Machine learning for automated text categorization,@ ACM CSUR,vol.34,no.1,pp.1-47,2002.
[10] Sun,Y.J.Han,J.Gao, and Y.Yu,"iTopicModel: Information network integrated topic modelling," in Proc. ICDM Conf., New York,NY,USA,1997,pp.109-110.
[11] Xu.W,X,Liu, and Y.Gong ,@Document clustering based on non-negative, matrix factorization," in Proc. ACM SIGIR Conf.,New York,NY,USA,2003,pp.267-273.
[12] Yang.Y, Y. Cao, Z. Nie, J. Zhou, and J. Wen, "Closing the Loop in Webpage Understanding," IEEE Trans. Knowledge and Data Eng., vol. 22, no. 5, pp. 639-650, May 2010.
[13] Zhong.S,"Efficient streaming text clustering," Neural Netw.,vol.18,no.5-6,pp.790-798,2005.
[14] Zhang.T ,R.Ramakrishnan and M.Livny ,"BIRCH: An Efficient data clustering method for very large databases," in Proc. ACM SIGMOD Conf, New York, NY,USA,1996,pp:103-114.
[15] Zhao.Y and G.Karypis,"Topic-driven clustering for document datastes," in Proc.SIAM Conf .Data Mining, 2005, pp.358-369.
[16] Zhou.Y , H.Cheng and J.X.Yu, "Graph clustering based on structural/attribute similarities, "PVLDB, vol.2, no.1, pp.718-729,2009.