Research Paper
Impact Factor: 3.567
Peer Reviewed Journal

IJMDRR
E- ISSN –2395-1885
ISSN -2395-1877

WEB NAVIGATION DATA MINING

Pradeep Kumar Shriwas* **Deepesh Dewangan****
*(M.Tech, Object Oriented Software Development), Department of Computer Science, Kalinga University, Raipur.
**Assistant Professor, Department of Computer Science, Kalinga University, Raipur.

I. Abstract

There is a huge amount of data available in the Information Industry. This data is of no use until it is converted into useful information. It is necessary to analyze this huge amount of data and extract useful information from it. Extraction of information is not the only process we need to perform; data mining also involves other processes such as Data Cleaning, Data Integration, Data Transformation, Data Mining, Pattern Evaluation and Data Presentation. We would be able to use this information in many applications such as Fraud Detection, Market Analysis, Production Control, and Science Exploration. The ***objective*** *of this paper is to introduce the different methodology used in web mining. What are the advantage and disadvantage of different methods associated in data mining. Finding which is the best method to mine the data in which area with the reference of these we can do the research on new method or modified technique in web data mining?*

We will analyse the: Apriori Algorithms, TSP Algorithm, Path Traversal Algorithm, FP Growth Algorithms and their impact in web navigation data mining.

Key Word: Apriori Algorithm, Path Traversal, Frequent Pattern, TSP, Web Navigation, Data Mining.

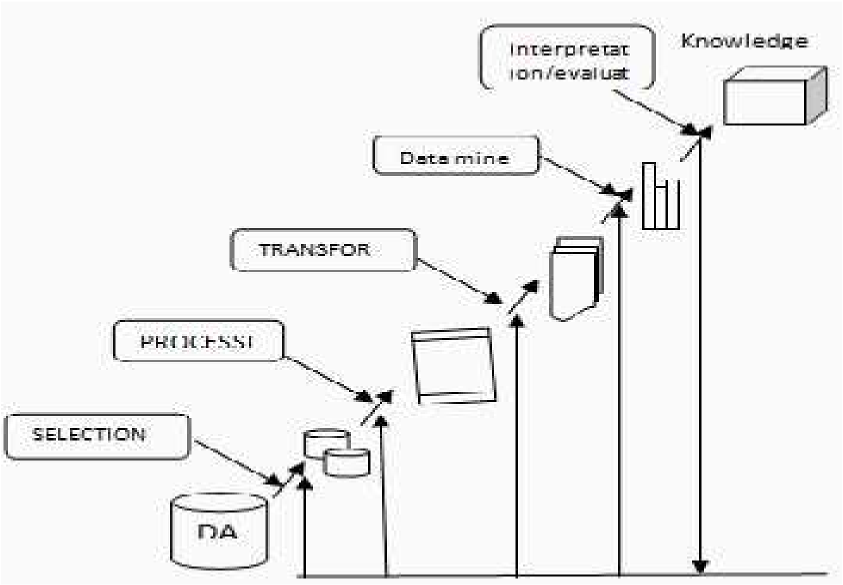
II. Introduction

The term Knowledge Discovery in Databases, or KDD for short, refers to the broad process of finding knowledge in data, and emphasizes the "high-level" application of particular data mining methods. It is of interest to researchers in machine learning pattern recognition, databases, statistics, artificial intelligence, knowledge acquisition for expert systems, and data visualization.

The unifying goal of the KDD process is to extract knowledge from data in the context of large databases.

It does this by using data mining methods (algorithms) to extract (identify) what is deemed knowledge, according to the specifications of measures and thresholds, using a database along with any required pre-processing, sub sampling, and transformations of that database.

A mining as an essential step in the process of knowledge discovery. Here is the list of steps involved in the knowledge discovery process:



The database of web is in the form of web sessions with session id or session number, this type of mining we called Web Mining. And in the Web Mining when we use to find path traversal pattern for decision management in website design, then it comes under the web usage Mining.

III. Literature Survey

Effective Web Log Mining and Online Navigational Pattern Prediction
Author: Abdelghani Guerbas (Department of Computer Science, University of Calgary, Calgary, Alberta, Canada)

Abstract: Accurate web log mining results and efficient online navigational pattern prediction are undeniably crucial for tuning up websites and consequently helping in visitors' retention. Like any other data mining task, web log mining starts with data cleaning and preparation and it ends up discovering some hidden knowledge which cannot be extracted using conventional methods. In order for this process to yield good results it has to rely on some good quality input data. Therefore, more focus in this process should be on data cleaning and pre-processing. On the other hand, one of the challenges facing online prediction is scalability.

Conclusion: In this paper, we proposed an efficient framework for web log mining and for online navigational behaviour prediction. We have reviewed all steps of this process and we have analyzed existing approaches and made an effort to make a contribution at each step. The first step in web log mining consists of different phases. The first phase is data cleaning and preparation. Enter a second phase of session identification as clarified .The main problem at this level is how the sessions' identification must be done. By checking the web log we can tell when a visitor had started browsing a website but it is very hard to tell when the visitor had left. Several approaches to identify sessions such as time-out based techniques have been proposed by researchers and have been used in commercial products.

Hybrid Technique for User's Web Page Access Prediction Based on Markov Model
Author: Prof. Urmi D. Agravat (Department of IT, A.D.Patel Institute of Technology)

Abstract: Web Mining consists of three different categories, namely Web Content Mining, Web Structure Mining, and Web Usage Mining (is the

process of discovering knowledge from the interaction generated by the users in the form of access logs, browser logs, proxy-server logs, user session data, cookies). This paper present mining process of web server log files in order to extract usage patterns to web link prediction with the help of proposed Markov Model.

Conclusions: We considered the problem of predicting user's access to the web server. Tests conducted on the Markov models built, using the logs data from the web access server of www.adit.ac.in organizational institute. Accurately predicting Web user access behaviour can minimize user-perceived latency, which is crucial in the rapidly growing World Wide Web. Although traditional Markov models have helped predict user access behaviour, they have serious limitations. So that, the novel clustering markov model with apriori algorithm presented, analysed and evaluated to predict Web access precisely which providing high accuracy.

Novel Technique for Mining Closed Frequent Item Sets Using Variable Sliding Window
Author: Vikas kumar (Department of CSE, Garg engineering college Ghaziabad, India)

Abstract: Frequent item set mining over dynamic data is an important problem in the context of knowledge discovery and data mining. Various data stream models are being used for mining frequent item sets. In a data stream model the data arrive at high speed such that the algorithms used for mining data streams must process them in strict constraint of time and space. Due to emphasis over recent data and its bounded memory requirement, sliding window model is a widely used model for mining frequent item set over data stream. In this paper we proposed an algorithm named *Variable- Moment* for mining both frequent and closed frequent item set over data stream.

Conclusion

Considering the continuousness of a data stream, the traditional methods or techniques for finding frequent item sets in conventional data mining methodology may not be valid in a data stream. This is because we cannot consider whole data and must identify when a data becomes obsolete or invalid. As the old information of a data stream may be no longer useful or possibly invalid at present. In order to support various requirements of mining data stream, the mining window or the interesting recent range of a data stream needs not to be defined static but must be flexible. Based on this range, a data mining method can be able to identify when a transactions becomes stale or needs to be disregarded.

Web Usage Mining Using Improved Frequent Pattern Tree Algorithms
Author: Ashika Gupta (IITM Gwalior (MP)-INDIA)

Abstract: Web mining can be broadly defined as discovery and analysis of useful information from the World Wide Web. Web Usage Mining can be described as the discovery and analysis of user accessibility pattern, during the mining of log files and associated data from a particular Web site, in order to realize and better serve the needs of Web-based applications. Web usage mining itself can be categorised further depending on the kind of usage data considered they are web server, application server and application level data. This Research work focuses on web use mining and specifically keeps tabs on running across the web utilization examples of sites from the server log records.

Conclusion: Web usage mining is the procedure of finding out which users are looking for the internet. It can be described as the sighting and scrutiny of user ease of access pattern, during mining of files and its connected data from a Web site, in order to recognize and better serve up the desires of Web-based applications.

One of the algorithms which is very simple to use and easy to implement is the Apriori algorithm.

Frequent Pattern Mining Using Semantic FP-Growth for Effective Web Service Ranking
Author: Omair Shafiq (Department of Comp. Sc, University of Alberta, Canada)

Abstract: Automated Ranking is crucial in the process of automated Web Services execution. Often adaptation and ranking (used interchangeably) of discovered Web services is carried out using functional and non-functional information of Web Services. Existing approaches are either found to be only focusing on semantic modelling and representation only, or using data mining and machine learning based approaches on unstructured and raw data to perform discovery and ranking. We propose an approach to allow semantically formalized representation of logs during Web Service execution and then use such logs to perform ranking and adaptation of discovered Web Services. We have built Semantic FP-Tree based technique to perform association rule learning on functional and non-functional characteristics of Web Services.

Conclusions: In this paper, we proposed a unique approach for ranking and adaptation of Web Services using Association Rule Mining based on our proposed Semantic Logs and Semantic extension of FP-Growth. Our proposed approach allows semantically formalized representation of logs during Web Service execution which are then used to perform ranking and adaptation of the discovered Web Services.

IV. Methodology

1. Apriori Algorithms

Apriori is an algorithm for frequent item set mining and association rule learning over transactional databases. It proceeds by identifying the frequent individual items in the database and extending them to larger and larger item sets as long as those item sets appear sufficiently often in the database. The frequent item sets determined by Apriori can be used to determine association rules which highlight general trends in the database: this has applications in domains such as market basket analysis.

2. Path Traversal Algorithm

It is the set of vertices where each vertex shows a web-page which are connected with each other by using link and via-link. In the implementation of path traversal graph, links and via- link are associated with vertices which represent the web page. As per traversal time, each web page appears in traversal sequence and there is chance of repetition of web-page.

3. FP-Growth Algorithm

FP stands for frequent pattern. In the first pass, the algorithm counts occurrence of items (attribute-value pairs) in the dataset, and stores them to 'header table'. In the second pass, it builds the FP-tree structure by inserting instances. Items in each instance have to be sorted by descending order of their frequency in the dataset, so that the tree can be processed quickly. Items in each instance that do not meet minimum coverage threshold are discarded.

If many instances share most frequent items, FP-tree provides high compression close to tree root.

4. TSP Algorithm

The TSP finds proves more effective to predict one step forward visit to next web page. As we get predictions of visitor, website operator can efficiently reconfigure the personalized website structure and take the help of throughout-surfing patterns for rearranging the contents of the website. First we collect the user login and surfing session from the web server. The surfing session contains the session id and consecutive browsing sessions. Each horizontal line in a session id shows the access pages of a particular website visitor. We draw the frequent path traversal graph which will show all possible connection for each web page by using a dataset of web browsing session. In the graph from to via-link information is important for user's to predict where a visitor will go at any vertex by the vertex he comes from. In proposing modified algorithm we will skip the TSP, which avoids the repetitive database scan of data sessions and also neglect unwanted and duplicate data. Result of the filtering module contain user navigation pattern in the form of hyper links.

V. Problem Definition

Apriori Algorithm of Mining Association Rules

1. Assume transaction database is memory resident.
2. Require many database scans.
3. Apriori, while historically significant, suffers from a number of inefficiencies or trade-offs, which have spawned other algorithms.

Candidate generation generates large numbers of subsets (the algorithm attempts to load up the candidate set with as many as possible before each scan).

Bottom-up subset exploration (essentially a breadth-first traversal of the subset lattice) finds any maximal subset S only after all $2^{|S|} - 1$ of its proper subsets.

Path Traversal Algorithm

1. Since user’s backward browsing is treated as for ease of traveling, there exist a possibility that some information about the user’s browsing behavior get lost. It is better that the backward traversal pattern can also be mined to reflect the user’s real behavior.
2. When traversal log record only contains destination references instead of a pair of references. The **MF** algorithm cannot identify the breakpoint where the user picks a new URL to begin a new traversal path. This could increase the computational complexity because the paths considered become longer, especially in an environment where users jump from sites to sites frequently. Thus some complement method should be employed to deal with this problem.
3. Problem related to pattern prediction.
4. Problem related to log cleaning.
5. Problem related to log acquisition of data.

VI. Comparative Result

Parameters	Apriori Algorithm	FP-Growth Algorithm
Technique	Use Apriori Property and join and prune property	It constructs conditional frequent pattern tree and conditional pattern base from database which satisfy minimum support.
Memory Utilization	Due to large Number of candidate are generated so require large memory space.	Due to compact structure and no candidate generation require less memory.
No. of Scans	Multiple scans for generating candidate sets.	Scan the database twice only.
Time	Execution time is more as time is wasted in producing candidates every time.	Execution time is small than Apriori algorithm,

VII. Conclusion

1. FP-Tree: a novel data structure storing compressed, crucial information about frequent patterns, Compact yet complete for frequent pattern mining.
2. FP-Growth: an efficient mining method of frequent pattern in large database: using highly compact FP-Tree, divide – and – conquer method in nature.
3. Both Apriori and FP – Growth are aiming to find out complete set of patterns but FP-Growth is more efficient than Apriori in respect to long patterns.
4. TSP with FP Growth - For graphical representation of a website the directed graph contains edges and vertices corresponding to hyperlinks and documents respectively. By mining the throughout-surfing patterns, we can predict the next node to be visited, so there is need to know from where the visitor comes from. For that "from-to-via" link concept is used as we discussed already in introduction. Also this via-link concept is mainly used for mining of TSP. Therefore novel data structure called Frequent Pattern Growth Algorithm is used that consisting set of vertices, edges, and via-links to store the information from web browsing sessions. The nodes traversed in graph are put in traced and untraced stack which will help us to form different patterns as per the visitor's interest.

VIII. Future Work

We are going to propose a framework which is based on Dynamic Apriori Algorithm which will help to increase Precision, Recall, F- Measure, and Accuracy of the Pattern Determination by pre-processing, pattern discovery and users classification. Our work mainly focused on doing users classification on three bases: country based, site entry based and access time based classification. This helps in the efficient administration and personalization of the websites and for the increase profit.

References

1. Show-Jane Yen*, Yue-Shi Lee* and Min-Chi Hsieh, "An Efficient Incremental Algorithm for Mining Web Traversal Patterns", Proceedings of the 2005 IEEE International Conference on eBusiness Engineering (lCEBE'05).
2. Ming-Syan, Jong Soo, Philips Yu.,"Efficient Data Mining for Path Traversal Patterns", Proceeding of 1998 Knowledge and Data Engineering, IEEE Transactions on (Volume: 10, Issue: 2).
3. Ming-Yen Lin and Suh-Yin Lee, "Incremental Update on Sequential Patterns in Large Databases", Tools with Artificial Intelligence, 1998. Proceeding of Tenth IEEE International Conference.
4. R. Sri Kant, Rakesh Agrawal, "Mining Sequential Patterns: Generalization and Performance Improvements", IBM AJmaden Research Center (1996).
5. Show-Jane Yen and Arbee L.P. Chen "A Graph-Based Approach for Discovering Various Types of Association Rules", IEEE Transactions on Knowledge and Data Engineering, Vol. 13, No. 5, September/October 2001.
6. Anna Gutowska and Luis L. Perez "A Comparison of Methods for Classification and Prediction of Web Access Patterns", COMP540 - Final Report - Spring 20 I O.
7. Arthur.A.Shaw, N.P. Gopalan "Frequent Pattern Mining of Trajectory Coordinates using Apriori Algorithm", proceeding in International Journal of Computer Applications in volume 22-9, May 2011.