A Survey Paper on: Frequent Pattern Analysis Algorithm from the Web Log Data

Samiksha Kankane
Computer Science & Engineering
LNCT-E
Bhopal, India

Vikram Garg
Computer Science & Engineering
LNCT-E
Bhopal, India

ABSTRACT
Web data mining is an emerging research area where mining data is an important task and various algorithms has been proposed in order to solve the various issues related to the web mining in existing dataset. This paper focuses the concept of data mining and FP-Growth algorithm. As for FP-Growth algorithm, the effectiveness is limited by internal memory size because mining process is on the base of large tree-form data structure. This Research 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 procedure to work with the proposed technique which can be possible to remove the drawback of limitation of the existed technique in the web mining area.

The various web usages mining technique can further work on various scientific area, medical area and social media application to approach for the research and security related area. A detail and pattern growth technique can help in getting more data and further on using line up algorithm we can illustrate the data states presentation effectively.

Index Terms—Data Mining, Ranking, Clustering, Web Logs

1. INTRODUCTION
Web mining is the integration of information gathered by traditional data mining methodologies and techniques with information gathered over the World Wide Web. [2] It is used to understand customer behaviour, evaluate the effectiveness of a particular Web site, and help quantify the success of a marketing campaign. Content mining is used to examine data collected by search engines and web spiders. Structure mining is used to examine data related to the structure of a particular Web site and Web Usage Mining is applied to many real world problems to discover interesting user navigation patterns for Improvement of web site design by making additional topic or recommendations observing user or customer behaviour. They are web server data, application server data and application level data. Web server data correspond to the user logs that are collected at Web server.

Some of the typical data collected at Web server include IP addresses, page references, and access time of the users and is the main input to the present Research. This work concentrates on web usage mining and in particular focuses on discovering the web usage patterns of websites from the server log files.

1.1 Stages in Web Mining For Pattern Discovery

Data Preprocessing
The data should be preprocessed to improve the efficiency and ease of the mining process. The main task of data preprocessing is to prune noisy and irrelevant data, and to reduce data volume for the pattern discovery phase. Field Extraction and data cleaning algorithms parse the web log records separating the fields and purging.

Pattern discovery
Few techniques to discover patterns from preprocessed data are listed like converting IP addresses to domain names, filtering, dynamic site analysis, cookies, path analysis, association rules, sequential patterns, clustering, decision trees etc.

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

2. LITERATURE REVIEW
Hao Yan, Bo Zhang, Yibo Zhang, Fang -2010.In this paper author A WUM process extracts behavioral patterns from the Web usage data and, if available, from the Web site information (structure and content) and on the Web site users (user profiles). In this thesis, we bring two significant contributions for a Web Use Mining process. In paper author propose a customized application specific methodology for preprocessing the Web logs and a modified frequent pattern tree for the discovery of patterns efficiently.

Huiping Peng- 2010. In this paper the pattern analysis phase interesting knowledge is extracted from frequent patterns and these results are used for website modification. In this paper we are using the FP-growth algorithm for obtaining frequent access patterns from the web log data and providing valuable information about the user's interest.

Min Chen and young U. Ryu -2011. In this paper addresses how to improve a website without introducing substantial changes. Specifically, we propose a mathematical programming model to improve the user navigation on a website while minimizing alterations to its current structure. Results from extensive tests conducted on a publicly available real data set indicate that our model not only significantly improves the user navigation with very few changes, but also can be effectively solved.

Joy Shalom Sona, AshaAmbhaikar -2012.The research work present in this paper presents a comprehensive overview of web mining methods and techniques used for the evaluation of reconciling systems to achieve better web navigation efficiency in order to improve the efficiency of web site. It integrates and coordinates among different reasons for making recommendations including frequency of access, and patterns of access by visitors to the web site.

Table 1.0

Algorithms Storage	**Structure**	**Advantages**	**Disadvantages**
Apriori	Array based	Any subset of frequent item set is also frequent item set.	Multiple scans have to be done on database. Its time and Memory complexity is very large.
Apriori Tid	Array based	Number of entries may be smaller than the number of transactions in the data base.	The time and Space complexity is also very large
FP tree	Tree based	Scans the database only twice. It has interactive rule mining	It seems to be difficult in incremental. It require less memory and more execution time
Custom built Apriori	Array Based	Based on old Apriori algorithm. It is efficient and effective pattern analysis.	It requires more memory and more execution time.
FAP(frequent access pattern)	Tree Based	constructs frequent access pattern tree (FAP tree) as per the access paths and Next step is where the function of FAP-growth is used to mine both long and short access patterns on the FAP tree	No sequence among the elements in the data.

3. PROBLEM FORMULATION

Today the World Wide Web is popular and interactive medium to distribute information. The web is huge, diverse, dynamic and unstructured nature of web data, web data research encountered lot of challenges for web mining. Information user could encounter following challenges when interacting with web.

1. Finding Relevant Information- People either browse or use the search service when they want to find specific information on the web. Today's search tools have problems like low precision which is due to irrelevance of many of the search results. This results in a difficulty in finding the relevant information. Another problem is low recall which is due to inability to index all the information available on the web.

2. Creating new knowledge out of the information available on the web- This problem is basically sub problem of the above problem. Above problem is query triggered process (retrieval oriented) but this problem is data triggered process that presumes that already has collection of web data and extract potentially useful knowledge out of it.

3. Personalization of information- When people interact with the web they differ in the contents and presentations they prefer.

4. Learning about Consumers or individual users- This problem is about what the customer do and want. Inside this problem there are sub problem such as customizing the information to the intended consumers or even to personalize it to individual user, problem related to web site design and management and marketing

4. PROPOSED WORK

Here this paper proposed a technique on the basis of analysis of previously performed work on the mining web dataset.In order to find a relevant information about the frequent dataset here technique plan to perform the semantic and synaptic search on finding the correct information and thus it is going to perform the high level structure agent based semantic and synaptic search to find the usable dataset from the web such that it can be further usable for the web results and web research in web data mining technique.

SYNAPTIC WEB

A synaptic web mining is a technique which works on the branches of neuron based data searching and usage technique, a synaptic web mining demonstrate about the work associate with the link which are linked with the current associated link and further on, synaptic mining used, where a web search is required to get more precise result from the available web dataset.

The parameters technique consider as follows -

1. Recall.
2. Precision.
3. Accuracy.

5. METHODOLOGY OF WORK

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

A. Semantic Mining

A Web mining from the crawl is done first ,technique extracting the information from the web based on the similar type of object and their availability in semantic manner ,the data is been extracted and use to create Entropy.

B. Synaptic Mining

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

Lattice Construction: The basic element of the lattice is an atom i.e. single page. Each atom or page stands for length-1 prefix equivalence class. Beginning 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 observed from the various semantic data and user can optimize according to the visualisation.

Line-up is a technique which provide a procedure for the ranking optimization of data which provide the post ranking estimation and ranking using different attributes, which provide re-ranking of data using Line up procedure.

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

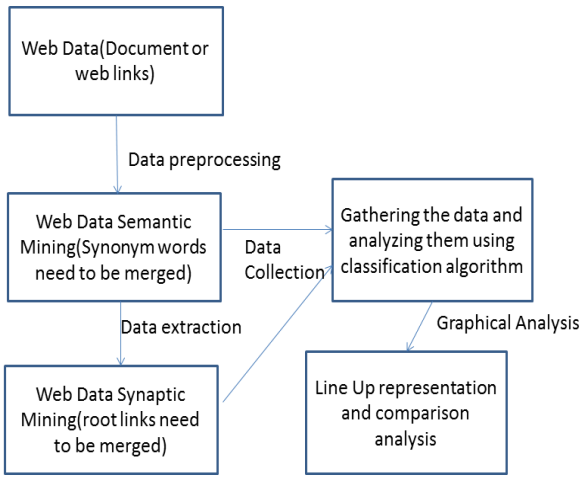

Fig-1.0(Proposed methodology flow)

6. EXPECTED OUTCOMES

This paper have analysed different schemes associated with the frequent item set mining in various web logs and information, here technique perform implementation and following results are expected, assumed to be monitored upon implementation of proposed work algorithm.

•Execution computation time: here this paper would like to monitor the execution time of process while processing the same dataset compare with the existing algorithm.

•Accuracy: Number of outcomes depends on the proposed and existing system should be compared with tabular and graphical format in the expected outcomes of the proposed work.

•Pattern number and computation should perform upon proposed work implementation.

7. CONCLUSION

The Approach followed by paper is new and two approach based algorithm which can process a data efficiently and it will help to find a best solution in the pattern recognition, The new approach requires minimum repeated database scan for frequent pattern mining in web usage mining. It will reduce the time and space execution.Web usage mining is the application of data mining techniques to discover usage patterns from Web data, in order to understand and better serve the needs of Web-based applications. In this paper a new technique is proposed to discover the web usage patterns of websites from the server log files with the foundation of clustering and improved Symantec. The effective algorithm will be proposed with the improvements as well as the implementation of Web mining approach and Apriori Algorithm. The forth coming step in the research work shall be to design the improved version of the Apriori Algorithm that shall be implemented on the Server Log Files for Association Rule Mining.

8. REFERENCES

[1] Hao Yan, Bo Zhang, Yibo Zhang, Fang Liu, Zhenming Lei "Web usage mining based on WAN users behaviours" 2010.

[2] Huiping Peng "Discovery of Interesting Association Rules Based on Web Usage Mining" 2010.

[3] Iqbal Gondal and Joarder Kamruzzaman Md. Mamunur Rashid, "Mining Associated Sensor Pattern for data stream oorks"Spain, 2013.

[4] Joy Shalom Sona, AshaAmbhaikar "Reconciling the Website Structure to Improve the Web Navigation Efficiency" July 2012.

[5] K. R. Suneetha, Dr. R. Krishnamoorthi, "Identifying User Behaviour by Analyzing Web Server Access log"2009.

[6] Luca Cagliero and Paolo Garza "Infrequent Weighted Itemset Mining using Frequent Pattern Growth", IEEE Transactions on Knowledge and Data Engineering, 2013.

[7] Min Chen and young U. Ryu "Facilitating Effective User Navigation through Website Structure Improvement" IEEE KDD, 2011.

[8] Sanjay Kumar Malik, Nupur Prakash, SamRizvi" Ontology and Web Usage Mining towards an Intelligent Web focusing web logs" 2010.

[9] Xiaoting Wei, Yunlong, Feng Zhang, Min Liu, WeimingShen Incremental FP-Growth Mining Strategy for Dynamic Threshold Value and Database Based on MapReduce School of Electronic and Information Engineering, Tongji University, Shanghai ,China IEEE2014.

[10] Samuel Gratzl, Alexander Lex, Nils Gehlenborg, Hanspeter Pfister and Marc Streit, "LineUp: Visual Analysis of Multi-Attribute Rankings" IEEE 2013.