An Enhanced Frequent Pattern Analysis Technique from the Web Log Data

[1]Iqbaldeep Kaur, [2]Navneet Kaur, [3]Nafiza Mann, [4]Isha Vats
[1,2,3,4]Dept. of CSE, Chandigarh Engineering College, Landran, Punjab, India

Abstract

To improve user experience while accessing the, website. Web usage mining is used to evaluate user's previous experiences, which helps to improve functionality of that website. In this paper a technique for web usage mining is proposed, which extends features of synaptic search and Frequent Pattern Growth algorithm. Proposed technique uses synaptic search property to search data on web on the basis of location and uses FP growth algorithm to generate results.

Keywords

Data Mining, FP Growth, Synaptic Search, Semantic Search, Web Logs.

I. 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.

A. Stages in Web Mining For Pattern Discovery

1. 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.

2. 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.

3. 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.

II. Literature Review

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

HuipingPeng - 2010. In this paper the interesting knowledge isextracted from frequent patterns and these results are used for website modification. In this paper the FP-growth algorithm is used 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. This paper addresses how to improve a website without introducing substantial changes. Specifically a mathematical programming model is used 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 Thispaper presentsa 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 website.

III. 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.

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.

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.

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

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

IV. Proposed Work

This paper proposed a technique on the basis of analysis of previously performed work for mining web dataset. In order to find a relevant information about the frequent dataset. Here technique plans to perform the semantic and synaptic search for finding the correct information. 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.

A. 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 -

- Recall:-Recall shows that value fetched by the algorithm is relevant to the query or most of the values are relevant to the query.
- Precision:-It is also known as positive predictive values. It is the fraction of the relevant values that are retrieved from the data set. This technique returns most relevant results as compare to irrelevant results.
- Accuracy: Accuracy is the measure of the correctness of the values that retrieved from the database in context of the query.

V. Methodology of Work

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

Lattice Construction: The basic element of the lattice is anatom 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 provides 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.

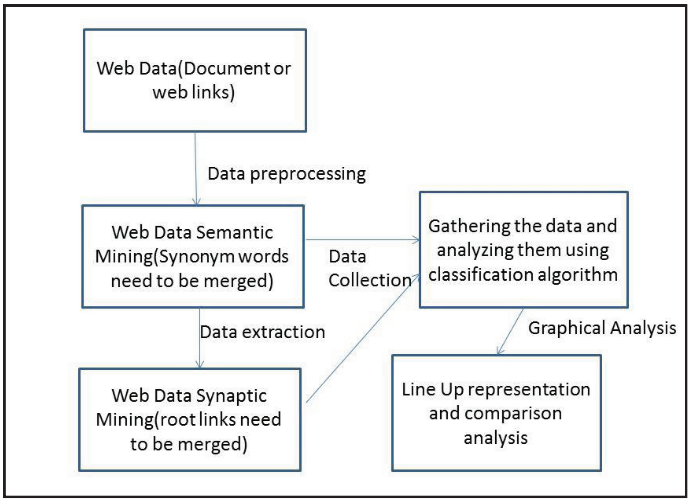
Web Data(Document or web links)
Data preprocessing
Web Data Semantic Mining(Synonym words need to be merged)
Data Collection
Gathering the data and analyzing them using classification algorithm
Data extraction
Graphical Analysis
Web Data Synaptic Mining(root links need to be merged)
Line Up representation and comparison analysis

Fig. 5: Proposed Methodology

VI. Experiment Results and Analysis

To implement proposed technique, Java language is used over Net Beans Ide and live database of www.freshersworld.com is use as dataset. In proposed work, first data is searched in database of freshersworld.com and provides normal search result. It will give semantic search result on the basis of related objects. It will provide synaptic search result and trace the location from normal search and generate result. It will provide whole dataset and split it in individual dataset and trace frequent &infrequent items, consider count as 2.Find support & count. Finally it will show FP growth result according to user visualization. It take Accuracy and Time to measure the performance of the technique and compare it with existing technique.

Accuracy: - A comparative analysis curve is shown below which shows that proposed technique provides 100% accurate results.

In this technique, 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.

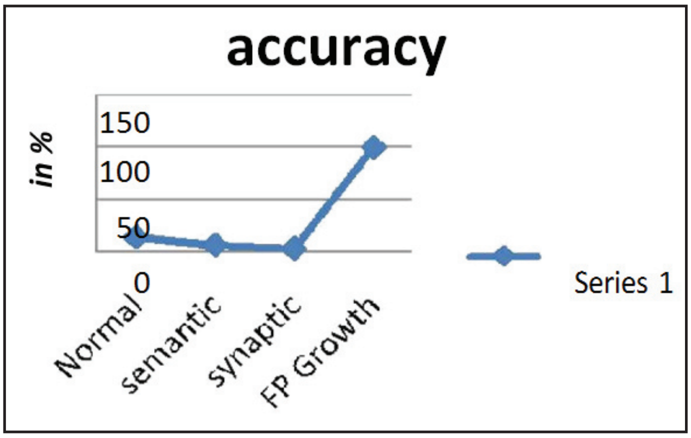

Fig. 6: Accuracy curve

Execution time: - In below fig. 2, a curve for execution time analysis is shown which shows a comparative analysis of execution time among the techniques.

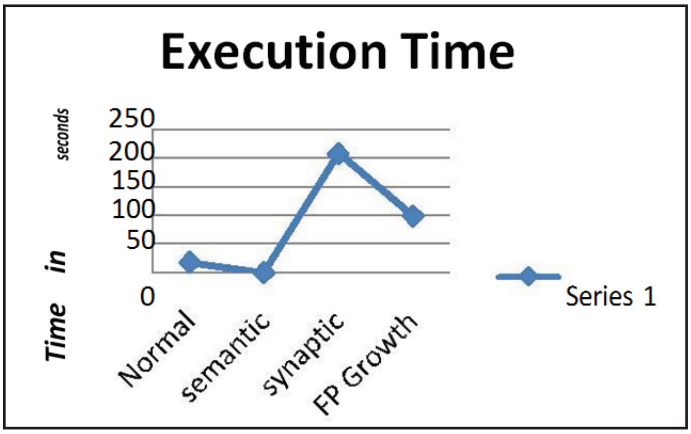

Fig. 6: Execution Time Curve

7. Conclusion and Future Work

A. Conclusion

There are many search techniques that are used to search data on web or fetch web usage mining. In this paper a technique based on FP Growth algorithm for web usage mining is proposed. A comparative analysis is shown in the figure 6.1, which shows that proposed technique provides most accurate result compared to other existing techniques. A time analysis in figure 6.2 is also presented which shows that proposed technique provides most accurate results in a small time span. Thus this technique is efficient to provide accurate results for web usage mining in small time span.

B. Future Work

The Combination of web mining techniques with technology will lead to improved performance, reduced network traffic and better results. Enhancement for such is still required. Work presented in this paper there still a scope for enhancement in time to provide fast result for search query. That can enhance the performance of the technique.

References

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

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

[3] IqbalGondal 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] Rahul Mishra, AbhaChoubey "Discovery of Frequent Patterns From Web Log Data by using FP growth Algorithm for web usage mining" 2012.

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

[10] Sanjay Kumar Malik, NupurPrakash, SamRizvi" Ontologyand Web Usage Mining towards an Intelligent Web focusing web logs" 2010.

[11] Xiaoting Wei, Yunlong, Feng Zhang, Min Liu, WeimingShen Incremental FP-Growth Mining Strategy for Dynamic Threshold Value and Database Based on Map Reduce School of Electronic and Information Engineering, Tongji University, Shanghai ,China IEEE201.

[12] Amit Verma et al.,"Optimization for Energy Efficient Mobile Device Protocols Stack Over Wireless Sensors", National Conference on Challenges in Emerging Computer Technologies (CECT-2010), held at Rayat Bahra Institute of Engineering and Bio Technology, Vol. 1, pp. 429-432, 8th-9th April, 2010.

[13] Amit Verma et al.,"Erbium Doped Fiber Amplifiers For Wavelength Multiplexing Systems", National Conference Cum Workshop on Information Security and Networks (ISAN-2009), pp. 94-96, Held at CIET-RAJPURA-PUNJAB on 19th -20th, June 2009.