Survey On: Web Usage Mining Techniques Generating Frequent Patterns for Web-Page Recommendation.

Gauri A Sonawane[1], Prof.Dr.Mrs.S.A.Itkar[2]

Dept.Computer Engineering, PES Modern College Of Engineering
Savitribai Phule Pune University
Pune,India

***Abstract*— Web page recommendation has become popular and efficient in the world of web. As most of the website recommend web pages according to the need of the user proper recommendation is an important task. Many techniques are present which help in web page recommendation out of which is finding effective usage patterns from the web log data. Useful knowledge discovery from Web usage data and satisfactory knowledge representation for effective Web-page recommendations are crucial and challenging. This paper shows a survey of various usage mining techniques which can be used to effectively generate frequent which gives the Web usage knowledge of a website.**

***Keywords*— Web Usage Mining (WUM),Web Page Recommendation.**

I. Introduction

The growing information on the World Wide Web with the development of advanced electronic devices has made Web information increasingly important. The developing introduction of current websites has overwhelmed Web users by offering many choices. But, web users don't make good decisions when surfing the web due to an inability to cope with enormous amounts of information [1].Recommender systems have proved in recent years to be a valuable means of helping web users by providing useful and effective recommendations or suggestions. Web-page recommendation has become immensely well known, and is shown as links to relevant stories, relevant books, or most viewed pages at websites. When a user searches a website, a sequence of visited Web-pages during a session (the period from starting, to existing the browser by the user) can be created. This sequence is arranged into a Web session $S = d_1d_2 \ldots d_k$, where d_i ($i = [1 \ldots k]$) is the page ID of the i^{th} visited Web-page by the user. The goal of a Web-page direction system is to efficiently forecast the Web-page or pages that will be visited from a given Web-page of a website. There are a number of problems in building an efficient Web-page direction system, such as how to efficiently learn from available historical information and search important knowledge of the field and Web-page navigation patterns, how to model and use the discovered knowledge, and how to make efficient Web-page directions based on the discovered comprehension. A great deal of research has been devoted to resolve these problems over the past decade. It has been reported that the approaches depends on tree structure and probabilistic models can effectively demonstrate Web access sequences (WAS) in the Web usage data [2].These approaches learn from the training datasets to construct the transition links between Web-pages. By implementing these approaches, given the existing visited Web-page (referred to as a state) and k previously visited pages (the previous k states), the Web-page(s) that will be visited in the next navigation step can be forecasted. The operation of these approaches based on the sizes of training datasets. The bigger the training dataset size is, the higher the prediction accuracy is. However, these approaches make Web-page direction solely based on the Web access sequences learnt from the Web usage data.

II. System Flow

The system flow shows that the web logs are pre-processed and the required data is passed to the web usage mining phase which generates the frequent sequences used for recommendation.

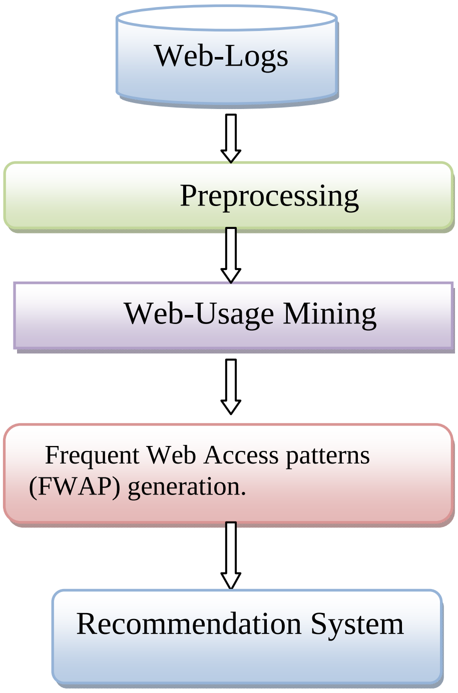

Fig.1 System Flow

1) Weblog: The raw data is in a web log, recording information about the surfing histories of the Web users.
2) Web Mining Techniques: An advanced Web usage mining techniques are used to discover the Web usage knowledge, which is in the form of frequent Web access patterns (FWAP), i.e patterns of frequently visited Web-pages.
3) Recommendation Engine: It will actually recommend the web pages to the user.

III. Literature Survey

A. Web Usage Mining.

Web Usage Mining (WUM) is the application of data mining techniques to discover interesting usage patterns from Web data, in order to understand and better serve the needs of Web-based applications. Usage data captures the identity or origin of Web users along with their browsing behaviour at a Web site. Web usage mining itself can be classified further depending on the kind of usage data considered. 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 a Web server include IP addresses, page references, and access time of the users and is the main input to the present Research. This Research work concentrates on web usage mining and in particular focuses on discovering the web usage patterns of websites from the server log files.

B. Algorithms

The algorithms for mining the frequent patterns can be categorized as follows:

1) Frequent Itemset Mining Algorithms.

i) Apriori-Apriori comes under frequent itemset mining algorithm as it is used to mine all frequent itemsets in database .It makes many searches in database to find frequent itemsets where k-itemsets are used to generate k+1-itemsets. Each k-itemset must be greater than or equal to a threshold value called minimum support threshold to be frequent. Candidate itemsets are generated at every step.Apriori works on two steps the *prune step* and the *join step* In the first, the algorithm scan database to find frequency of 1-itemsets that contains only one item by counting each item in database. The frequency of 1-itemsets is used to find the itemsets in 2- itemsets which in turn is used to find 3-itemsets and so on until there are not any more k-itemsets. If an itemset is not frequent, any large subset from it is also non-frequent this condition prune from search space in database.

2) Sequential Patterns Mining Algorithm.

ii)Generalized Sequential Patterns(GSP)-This algorithm works similar to the apriori algorithm as it also generates the candidate sets but in a different way. The first scan finds all of the frequent items which form the set of single item frequent sequences. Each subsequent pass starts with a *set* of sequential patterns, which is the set of sequential patterns found in the previous pass. This seed set is used to generate new potential patterns, called *candidate sequences*. Each candidate sequence contains one more item than a seed sequential pattern, where each element in the pattern may contain one or multiple items. The number of items in a sequence is called the *length* of the sequence. So, all the candidate sequences in a pass will have the same length. The scan of the database in one pass finds the support for each candidate sequence. All of the candidates whose support in the database is no less than min support form the set of the newly found sequential patterns. This set then becomes the seed set for the next pass. The algorithm terminates when no new sequential pattern is found in a pass, or no candidate sequence can be generated.

iii) PrefixSpan- This is a different algorithm from GSP and apriori as it does not generate any candidate set, instead uses a projected database to examine only the prefix subsequence's and project their corresponding postfix subsequence's into projected databases. In each projected database, sequential patterns are grown by exploring only local frequent patterns. To further improve mining efficiency, two kinds of database projections are explored: *level-by-level projection* and *bi-level projection*. Moreover, a main-memory-based pseudo-projection technique is developed for saving the cost of projection and speeding up processing when the projected (sub)-database and its associated pseudo-projection processing structure can fit in main memory.[9] PrefixSpan mines the complete set of patterns and is efficient and runs considerably faster than both based GSP algorithm and Apriori algorithm.

iv) Sequential Pattern Discovery using Equivalent Class (SPADE)-This algorithm utilizes combinational properties which decompose the original problem into smaller sub-problems. These problems can be independently solved in main-memory using efficient techniques mainly lattice search techniques, and simple join operations.[5] All sequences are discovered in only three database scans. Experiments show that SPADE outperforms the best previous algorithm by a factor of two, and by an order of magnitude with some pre-processed data. It also has linear scalability with respect to the number of input-sequences, and a number of other database parameters.

3) Pattern Growth Algorithms using tree structure

v) Web-access pattern(WAP-Mine)-This algorithm works using a tree structure where a tree is constructed. Items are considered as nodes. This algorithm can be mainly used when frequent sequences are to be found from the weblogs.The algorithm scans the web log two times once to find all frequent individual events and later to construct a tree over the set of frequent individual events of each transaction. It finds the suffix patterns[3].It constructs the intermediate conditional WAP-tree using the pattern found in previous step and mines the tree later to find the frequent sequences.

vi) Pre-ordered linked web access pattern(PLWAP-mine)-This is the sequential pattern mining algorithm which uses the tree structure. It constructs the tree taking the WASD which is a database of web access sequences and a minimum support value as input and generates FWAP as output [7].Initially the algorithm computes the frequent 1-itemset from each transaction. Using this frequent set it constructs PLWAP-tree by considering each item as a node. It labels the node in the format as (L:C:P) meaning Label,Count,Position code. The node which is the left child is assigned code 1, else if right child then 0. After construction of tree mining is performed by considering the header linkage of each node.

TABLE I
PROS AND CONS

Algorithm	**Pros**	**Cons**
Apriori	-Simple to implement. -Large itemset property.	Requires a lot of database scans.
GSP	Users can specify time constraint (Time between adjacent elements in a pattern).	Inefficient for mining long sequential patterns.
PrefixSpan	-Requires Less Projections. -No Candidate set generation.	Requires more time in construction of projected database.
SPADE	Requires less database scans.	Huge set of candidate generation.
WAP	Provides more scalability.	Constructs tree at every stage which may lead to time and memory wastage.
PLWAP	The process of mining the tree is simplified as it follows header linkage technique.	Scans the whole database which may be time consuming if the database keeps on increasing

IV. CONCLUSION

Web page recommendation is important and challenging task in the world of web. Proper recommendation is required which will take the user to their required web page. Many techniques and different ways are present to do this among which the web usage mining techniques are preferable by many researchers. These techniques work in a different way and have their own advantages and disadvantages. Selection of appropriate technique will give a better recommendation.

REFERENCES

[1] Konstan, J, Riedl, J,"Recommender Systems: from Algorithms to User Experience," *User Modeling and User-Adapted Interaction*, Springer 2012.

[2] Liu, B., Mobasher, B. & Nasraoui, O,"Web Usage Mining', in B. Liu (ed.), Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data", Springer-Verlag Berlin Heidelberg.

[3] Woon,Y.K., Ng, W.K. & Lim, E.-P,"Web Usage Mining: Algorithms and Results," in A. Scime (ed.), *Web Mining: Applications and Techniques*, IGI,2005.

[4] Khalil.F. "Combining Web Data Mining Techniques for Web Page Access Prediction," *2008 Doctoral thesis, University of Southern Queensland.*

[5] Mabroukeh, N.R. & Ezeife, C.I,"A Taxonomy of Sequential Pattern Mining Algorithms," ACM 2010.

[6] Henze, N., Dolog, P. & Nejdl, W, "Reasoning and Ontologies for Personalized E-Learning in the Semantic Web," *EducationalTechnology & Society*,2004.

[7] Ezeife, C. & Liu, Y. ,"Fast Incremental Mining of Web Sequential Patterns with PLWAP Tree," *Data Mining and Knowledge Discovery, vol. 19, no. 3.*Springer 2011

[8] Thi Thanh Sang Nguyen,Hay Yan Lu,Jie Lu"Web-PageRecommendation Based on Web Usage and Domain Knowledge,", *IEEE Transactions on knowledge and data engineering*, OCTOBER 2014.

[9] Jian Pei, Jiawei Han, Behzad Mortazavi-Asl, Helen Pinto" PrefixSpan: Mining Sequential Patterns Efficiently by Prefix-Projected Pattern Growth" *Natural Sciences and Engineering Research Council of Canada.*