A NOVEL FRAMEWORK ON WEB USAGE MINING

Ms. S. Sangavi[1], Mr. N. Senthil Kumaran[2]

[1]Research Scholar, Dept. of Computer Science, Vellalar College for Women, Erode - 12, India
[2]Asst. Professor and Head, Dept. of Computer Applications, Vellalar College for Women, Erode - 12, India

---***---

ABSTRACT: *Web Usage Mining (WUM) is the application of data mining techniques to discover interesting usable knowledge from Web log data. The goal of WUM is to retrieve, model, and analyze the access patterns of user with a Web site. WUM includes different phases of data mining techniques called Data Preprocessing, Pattern Discovery & Pattern Analysis. Initially, the web log is preprocessed to clean, integrate and transform into a common log. Later, Data mining techniques are applied to discover the interesting characteristics in the hidden patterns. The final stage of web usage mining is Pattern Analysis which can validate the interested patterns from the output of pattern discovery. The main objective of the proposed novel framework is to find the Sequential frequent pattern from the log file and predicting sequence rule from the frequent sequences. Many framework suggested on various pattern discovery where as the proposed work contain pattern discovery and pattern analysis phases. In the Pattern Analysis phase, the frequent sequential patterns are examined and Sequence Rule is predicted from the sequential pattern.*

Keywords: **Data Mining, Web Usage Mining, Pattern Discovery, Pattern Analysis, Sequential Frequent Pattern, Rule Prediction.**

I. INTRODUCTION

1.1 DATA MINING

Data mining or knowledge discovery in databases, as it known, is the non-trivial extraction of implicit, previously unknown and potentially useful information from the data. This encompasses a number of technical approaches, such as clustering, classification, finding dependency networks, analyzing changes and detecting anomalies.

Data mining is one of the advanced process or techniques to retrieve useful knowledge from vast collection of data. The derived knowledge from data base is also known as pattern or trend. Since the data base is very large, these patterns cannot be obtained by conventional data processing or searching.

1.2 WEB MINING:

Web Mining is the process of extracting knowledge from World Wide Web. It can be categorized as the data in actual web pages called web content mining, data in the structure of web site called structure mining, and data regarding the web activity called web usage mining.

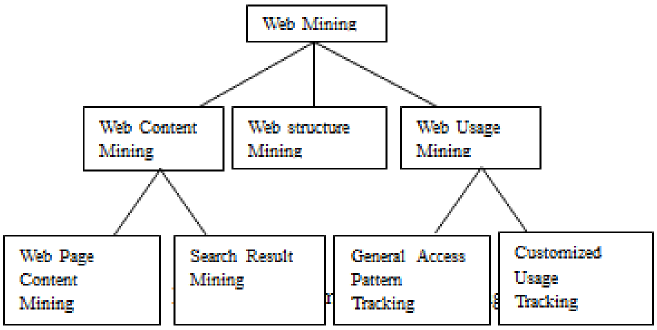

Fig 1.1 Taxonomy of Web Mining

1.2. WEB USAGE MINING:

Usage mining allows an organization to predict productive information contains the customer behaviour in online purchasing. The usage data that is gathered provides the companies with the ability to produce results more effective to their businesses and increasing of sales. Usage data can also be useful for improving online marketing skills that will be useful for the company to promote it's services on a higher level. Usage mining is useful not only in online marketing, but also to e-businesses, e-commerce applications and commercial organization depends on online registration. Web usage mining is used to give recommendation for the end user to select web pages to reach the correct pages.

1.2.2 PHASES OF WEB USAGE MINING:

WUM consists of three independent phases called Data collection and Pre-processing, Pattern Discover and Pattern Analysis.

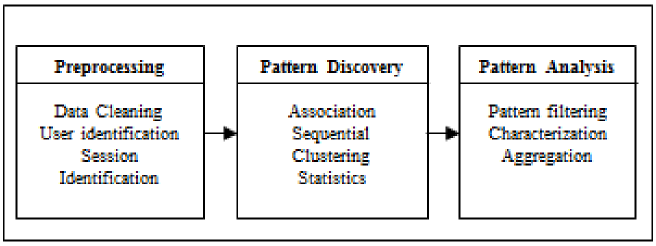

Fig 1.2.1 Phases of Web Usage Mining

Pre-Processing

An important task in WUM is to create and prepare a data set for suitable mining processes. The creation of a suitable target data set is crucial task. In Web usage mining process the data preparation process is the most time consuming and computationally intensive step.

Pattern Discovery

The preprocessed data is considered for the knowledge extraction algorithm such as AI, Data Mining algorithms, and information theory. Different data mining algorithms like path analysis, association rule, sequential pattern, clustering and classification are used for effective process of data mining. When exposed to these algorithms, data in the web access logs can be transformed into knowledge to uncover the potentials in the pre-processed data, which further requires analysis.

Pattern Analysis

The last phase in the web usage mining is pattern analysis of the obtained results in order to distinguish trivial, useless knowledge from knowledge that could be used for website modification, system improvement and web personalization. The common technique used for pattern analysis are, visualization technique, OLAP technique, Data and knowledge Querying and usability analysis.

II RELATED WORK

R. Agrawal, et al., [1995], presented a new algorithm of mining sequential patterns from a database of customer sales transactions and presented three algorithms for solving this problem. Two of the algorithms, AprioriSome and AprioriAll, have comparable performance, although AprioriSome performs a little better for the lower values of the minimum number of customers that must support a sequential pattern. Experiments show that both AprioriAll and AprioriSome shows similar performance based on customer transactions.

S.Parthasarathy, et al., [1999] proposed Incremental and interactive sequence mining ISM: which deals with incremental sequence mining for vertical database based on the SPADE approach of sequential pattern mining. Additionally, ISM prunes the search space for potential new sequences based on the construction of Incremental Sequence Lattice (ISL) and the exploration of its properties. Performance study shows that ISM is an improvement in execution time up to several orders of magnitude in practice, both for handling increments of the database, in addition to the handling interactive queries, compared with SPADE.

Dr. Sunita Mahajan, et al., [2014] proposed a work which implements sequential pattern mining algorithms on the Web sequential Datasets KDD CUP 2000 and MSNBC. As shown by the simulation results, Prefixspan gives best performance for dense dataset such as KDD cup 2000 and AprioriAll_Set an early pruning algorithm gives best performance for heavily dense dataset. Among the apriori based and pattern growth based algorithms SPADE performs best for heavily dense dataset but requires the maximum memory.

Manika Verma, et al.,[2014] made the comparison among GSP, SPADE and PrefixSpan algorithms. Present the concept and explanation of the above mentioned algorithms. Using SPMF it is determined that PrefixSpan is better than GSP and SPADE in performance as it takes lesser time than GSP and SPADE. SPADE takes less time than GSP but it takes quite more time than PrefixSpan.

Taral Patel, et al., [2014] discussed fundamental of sequential pattern mining and different sequential pattern mining algorithm. The main classification of sequential pattern mining algorithm is Apriori based and pattern growth based like GSP, SPADE, and SPAM. Also represent comparative study of these sequential patterns mining algorithm. Also discuss extension of sequential pattern mining algorithm like closed sequential pattern mining algorithm e.g. BIDE, Clospan and maximal sequential pattern mining algorithm e.g. MaxSP, MSPX, MFSPAN. These are given compact representation of sequential pattern mining algorithm.

Valli Mayil [2014] discussed the improved version of web usage mining called Ontology Based Semantic Web Usage Mining for Enhanced Recommendation Model. This model will enhance the preprocessing processes with semantic notations. For every web page the semantic nature of page is annotated. This will provide additional information for further processing. User identification and session identification process have sequence of information with semantic knowledge.

III METHODOLOGY:

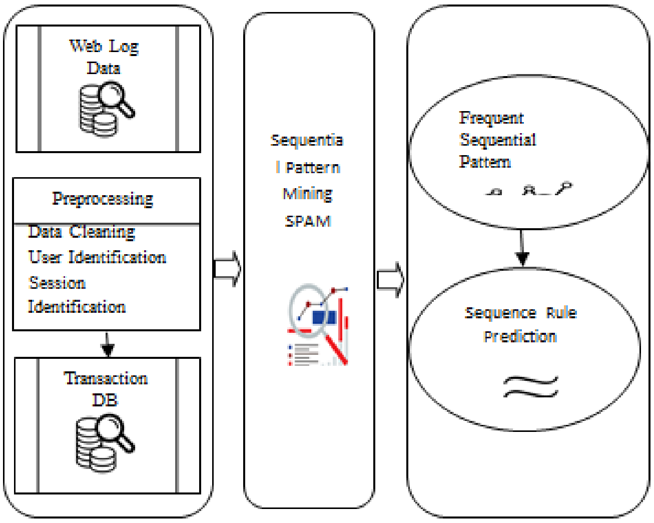

Fig 3.1 System Architecture

3.1 PREPROCESSING:

Data Cleaning: Data cleaning process involves such as, removing objects that may not be imperative for the purpose of analysis. It includes references to style files, graphics, or sound files. The cleaning process also may involve the removal of crawler in navigations.

User identification: The next task in preprocessing phase is user identification. Unique and individual user is identified to record their navigation patterns. Following are the methods to identify the user

- Authentication details
- Client side cookies
- IP address and Agent information

Session identification: Sessionization is the process of classifying the user activity record of each user into sessions. Group of pages visited by the user within specific time is considered as a session. Total time spent at the site cannot exceed the time limit is considered as a session.

Transaction database: Transaction data base is a temporary sequence data base contains sequence of session sequences**.** It is a sequence database consists of a set of n page views, P = {p1, p2, ⋯, pn}, and a set of 'm' user transactions, T = {t1,t2,⋯,tm}, where each ti in T is a subset of P. Page views are semantically meaningful entities to which mining tasks are applied (such as pages or products).

3.2 PATTERN DISCOVERY:

The process of sequential pattern mining attempts to find ordered set of pages visited by the customer. Applying SPAM algorithm would the marketers can predict future visit patterns which will be helpful in placing advertisements aimed at certain user groups. In the context of Web usage data, sequential pattern mining can be used to capture frequent navigational paths among user trails. SPAM discovers all frequent sequential patterns occurring in a sequence database that occurs in more than min sup sequences of the database. The support of a sequential pattern is the number of sequences where the pattern occurs divided by the total number of sequences in the database. A frequent sequential pattern is a sequential pattern having a support no less than the min sup parameter provided by the user.

3.3 PATTERN ANALYSIS – SEQUENCE RULE PREDICTION:

Sequence rule mining is a data mining technique to identify the new rule which will be used to make decision making.

A sequential rule is a rule of the form X -> Y where X and Y are sets of items (itemsets). A rule X ->Y is interpreted as if items in X occurs (in any order), then it will be followed by the items in Y (in any order). For example, consider the rule {a} -> {e,f}. It means that if a customer buy item "a", then the customer will later buy the items "e" and "f". But the order among items in {e,f} is not important. This means that a customer may buy "e" before "f" or "f" before "e".

To find sequential rules, two measures are generally used: the support and the confidence. The support of a rule X -> Y is how many sequences contains the items from X followed by the items from Y. The confidence of a rule X -> Y is the support of the rule divided by the number of sequences containing the items from X. It can be understood as the conditional probability P(X,Y|X).

IV RESULTS AND DISCUSSION:

In this section, the proposed framework is experimented and the performance is evaluated. The experiment of frequent Sequential Mining using SPAM and Sequence rule production system is implemented. In the data collection phase, MSNBC and Kosarak dataset of click-stream data have been collected from UCI repository.

The working of the proposed work is measured with metrics such as performance and scalability. In order to evaluate the performance of the proposed system, the metrics such as Memory Utilization of the algorithm is evaluated. The scalability of algorithm is evaluated by obtaining the measure such as maximum number of frequent pattern generated for various dataset.

Memory utilization with varying minimum support threshold

The performance of memory utilization of both SPAM and SPADE algorithm is compared and the results are obtained as follows in table 4.1.

Table 4.1 Memory utilization for the algorithms

Minimum support count in %	Memory consumed in MB for MSNBC		Memory consumed in MB for Kosarak	
	SPAM	SPADE	SPAM	SPADE
25	65	135	77	373
30	46	87	67	369
35	24.81	120	50	370
40	19.69	103	34	386
45	9.56	82	34	374
50	41	107	30	374

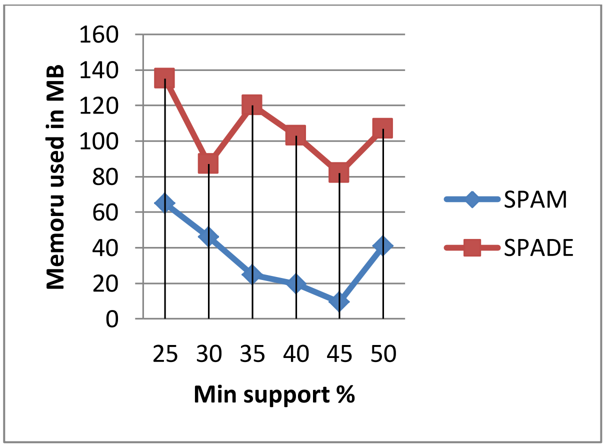

Figure 4.1(a) Memory used in MSNBC

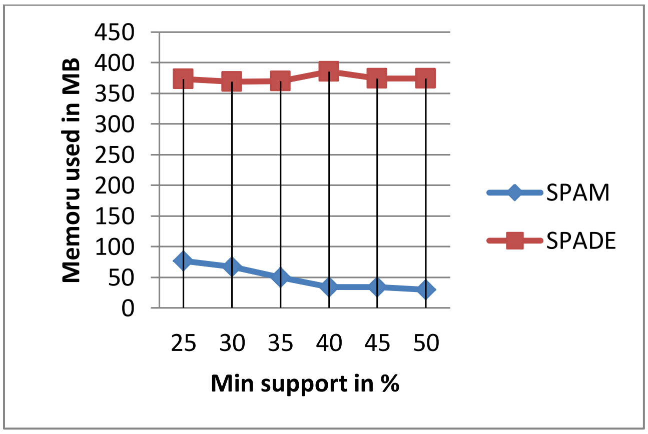

Figure 4.1 (b) Memory used in Kosarak

The graph in figure 4.1 shows the memory utilization of SPADE is larger compared to SPAM. This has been verified with different data set and the result is depicted in figure 4.1.

Scalability with number of frequent patterns generated with varying minimum support threshold for MSNBC

Scalability of algorithm is evaluated with the light and dense dataset for different patterns generated. Table 4.2 shows the result that number of sequences with different ranges.

Table 4.2 Number of patterns generated with different ranges

Minimum support count in %	Ranges in thousands	No. of sequences generated	Total patterns
25	<5	0	44
	>=5 to <10	20	
	>=10 to <15	16	
	>15	8	
30	<10	3	27
	10 to 15	16	
	15 to 25	8	

Sequence Rule Prediction with varying minimum support threshold for MSNBC Data set.

Min support in %	Number of sequence by SPAM	Sequence Rule pattern with min support

25	44	2==> 1 #SUP: 9 #CONF: 0.5625
		1== >1,1,1 #SUP: 9 #CONF:0.55
		1 ==> 2 #SUP: 10 #CONF: 0.3333333333333333
		4 ==> 1 #SUP: 2 #CONF: 0.25
		7 ==> 1 #SUP: 4 #CONF: 0.2
		7 ==>1,4 #SUP: 4 #CONF: 0.2
		12 ==> 1,2 #SUP: 2 #CONF: 0.5

30	27	2 ==> 1,1 #SUP: 5 #CONF: 0.5
		1 ==> 2,4 #SUP: 6 #CONF: 0.375

40	13	2 ==>2,3 #SUP: 2 #CONF: 0.3333333333333333
		1 ==> 2,4 #SUP: 3 #CONF: 0.375

V CONCLUSION AND FUTURE WORK:

WUM includes different phases of data mining techniques called Data Preprocessing, Pattern Discovery & Pattern Analysis. Initially, the web log is preprocessed to clean, integrate and transform into a common log. Later, Data mining techniques are applied to discover the interesting characteristics in the hidden patterns. Pattern Analysis is the final stage of web usage mining which can validate interested patterns from the output of pattern discovery.

The main objective of the proposed novel framework is to find the Sequential frequent pattern from the log file and predicting sequence rule from the frequent sequences. Many framework suggested on various pattern discovery where as the proposed work proposed pattern discovery and pattern analysis phases. In the pattern discovery phase, SPAM and SPADE algorithm is executed on MSNBC and kosarak dataset. The execution time and memory utilization is compared.

The future work of this thesis can be extended to include semantic notation in sequential pattern mining algorithm as well as in sequence rule prediction system. Semantic web search engine can be developed to easily locate the web pages in a recommendation system in various business and health application.

REFERENCES:

[1] R. Agrawal, R. Srikant , “Mining sequential patterns,” In Proceedings of Inter- national Conference on Data Engineering, pp. 3–14, 1995.

[2] Jiawei Han and Micheline Kamber, Data Mining - Concepts and Techniques, 3rd Edition, 2012, Morgan Kauffman Publishers.

[3] Manika Verma and Dr. Devarshi Mehta, “A Comparative study of Techniques in Data Mining”, International Journal of Emerging Technology and Advanced Engineering, Volume 4, Issue 4, April 2014, ISSN 2250-2459.

[4] Margaret H. Dunham, Data Mining Introductory and Advanced Topics, Prentice Hall, 2003.

[5] S. Parthasarathy, M. J. Zaki, M. Ogihara, S. Dwarkadas, ”Incremental and interactive sequence mining”, Proceedings of the eighth international conference on Information and knowledge management, Kansas City, Missouri, USA – November-1999, Pages 251-258.

[6] Dr. Sunita Mahajan, Prajakta Pawar and Alpa Reshamwala “Performance Analysis of Sequential Pattern Mining Algorithms on Large Dense Datasets”, International Journal of Application or Innovation in Engineering & Management (IJAIEM), Volume 3, Issue 2, February 2014.

[7] Taral Patel and Prof. Narendra Limbad,” A Study of Sequential Pattern Mining Algorithm”, IJSRD - International Journal for Scientific Research & Development, Vol. 2, Issue 09, 2014, ISSN (online): 2321-0613.