Volume 3, Issue 11, November 2013 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

A Survey on Web Usage Mining with Clicking Pattern in Grid Computing Environment

Archana Godbole
M.Tech (CTA) Research Scholar
RGTU University Bhopal, India

Mahendra Kumar Rai
M.Tech(IT) HOD
RGTU University Bhopal, India

Abstract— Web usage mining is a current and drastic research area in web usage mining focused on learning about web users and their interaction about web sites. The aim of web usage mining is to find user's access moves frequently and quickly from the massive web log data such as through frequent access paths, frequent access page groups and user clusters. Through web usage mining the whole registration information left by user access can be mined with the user access mode which provides foundation for decision making for big organizations. Web mining has become very crucial in those areas which are based upon web. So for managing massive web log data we need solid distributed environments like grids. Calculating user's browsing terms is a necessary operation of usage mining which generates exact usage data. In this paper we introduce a survey and analysis of latest web usage mining tricks and tactics in the form of algorithms with distributive grid environments for saving and using heavy data very easily from server in exact given time duration.

Keywords— Web Usage mining, pattern clicking algorithms, grid computing environment, Click streams

I. INTRODUCTION

A. Introduction to web usage mining

Today human life is totally depended on internet. On internet we can be searched anything and everything. Usually web pages contain vast amount of information so user may not interest on it, as it may not be the part of the main content of the web page. Web Usage Mining (WUM) is one of the main areas of data mining, artificial intelligence to the web data and forecast the user's visiting behaviours and obtains their interests through searching the samples. Since WUM explicitly indulge in some areas, like e-commerce, e-learning, Web analysis, information fetching etc. Weblog data is one of the main kind of resources which contain whole information related to the users visited pages, searching patterns, time spent on a particular link and this information can be used in various applications like adaptive web sites, updated services, customer summarization, pre-fetching, generate attractive web sites etc. There are lots of obstacles related with the existing web usage mining techniques. Practical applicability is major hurdle in existing web usage mining algorithms. This paper continues the line of research on Web access log analysis is to analyse the patterns of web site usage and the features of user habits. It is the truth that the normal Log data is very mushy and unclear thus it is important to pre-process the log data for effective web usage mining process. There are three phases in pre-processing which is user identification, data cleaning, and pattern discovery and pattern analysis. By nature web log data is noisy and unclear, thus for effective mining technique pre-processing is an efficient mining process. In this paper, a unique pre-processing trick is proposed by removing local and global noise and web robots. With the popularity of the internet, to understand the intentions of the internet users is of vital importance.web usage mining is part of the tasks trying to understand the users behaviours by analysing the web logs. its vary purpose is to discover useful, but originally hidden information or knowledge from massive web logs because web log data is increasing in seizes, one of the demands in web usage mining is the scalability ,i.e. the ability to handle massive data. Traditional algorithms are easily scalable, because their high algorithmic complexity and the limited capability to handle bigger datasets. Therefore, high performance parallel or distributed computation like grid environments provides a feasible solution to this problem.

B. Introduction to Grid Environment

In infrastructure the integrated and collaborative use of computer systems, networks and scientific instruments owned and managed by various organizations. Large amounts of data mainly include in Grid applications and/or computing resources that require secure resource sharing over organizational Boundaries. Thus this reason makes Grid application management and deployment a complex undertaking. Grid software provides users with joint computing ability and serial access to resources in the heterogeneous Grid Environment. Various software toolkits and systems have been developed, most of contributed in academic research projects, around all over the world. Some middleware like—UNICORE, Globus, Legion and Gridbus are used to provide flexible distributive grid environment for accommodation huge databases.

II. PROCESS OF WEB USAGE MINING

The Web Usage Mining is the process of applying techniques to detect patterns of usage to Web Page [3, 4]. The Web Usage Mining use the data storage in the Log files of Web server as first resource; Generally, three kinds of information

have to be handled in a web site: content, structure and log data. Content data contains of anything is in a web page, structure data is nothing but the organization of the content and usage data is nothing but the usage patterns of web sites. The usage of the data mining process to these dissimilar data sets is based on the three different research directions in the area of web mining like web content, structure and usage mining. Web usage mining also a data mining process for discovering the usage patterns from web information for the purpose of understanding and better provide the requirements of web-based applications. Web usage mining imitates the actions of humans as they interact with the Internet. Examination of user actions in communication with web site can offer insights causing to customization and personalization of a user's web practice. As a result of this, web usage mining is of extreme attention for e-marketing and ecommerce professionals. Namely three phases involves in web usage mining: pattern discovery, pre-processing, pattern analysis. There are different techniques available for web usage mining with its own advantages and disadvantages.

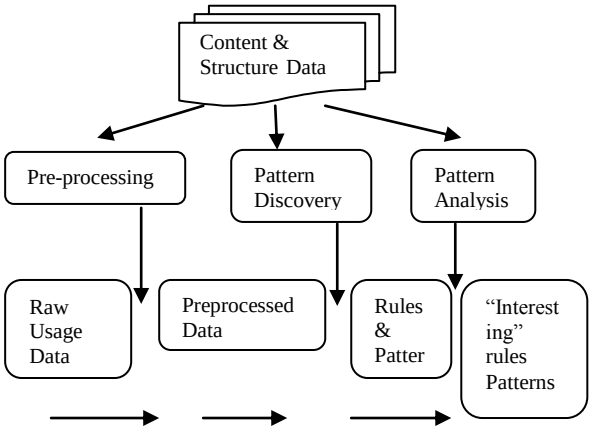

Fig 1. Process of WUM

This paper concerns the performance improvements of a sort of web usage mining task—the click stream analysis.

A. Conventional Methodologies of sequential clicking pattern Algorithms In Web usage Mining

This [6] section briefly discusses the conventional processes of sequential clicking pattern algorithms. Basically, the data sources for web usage mining are various web logs. Web logs are usually text files recording all transactions between the web server and the users [3]. Each transaction consists of a request of URL and a response and associated components (e.g. image files). Web logs may records for months of even for years, so that they are usually huge in size. Some search engines or web crawlers may invoke web spiders of robots to collect the information they need. Such activities are also recorded in the web logs. For the purpose of web usage mining for understanding the user's intentions, the following algorithms are used to mine useful knowledge in sequential clicking pattern manner.

- **Aprori Algorithm:** The Apriori algorithm was first proposed by Agrawal for the discovery of frequent item sets. It is the most widely used algorithm for the discovery of frequent itemsets and association rules. Any subset of a frequent itemset is a frequent itemset. Apriori is called a breadth first approach. the problem with the Apriori approach [8] is that it can generate a very large number of candidate sequences. For example, to generate a Quick sequence of length 100, it should generate 2100 candidates, because these are its sub sequences, which are also frequent.

- **SPADE (Sequential Pattern Discovery using Equivalent Class):** In [9] Zaki proposes an algorithm that takes a different approach than Apriori and GSP for the discovery of sequential patterns. Instead of assuming a horizontal database, where a customer has a set of transactions associated with him, SPADE assumes a vertical database, where each item is associated with a customer id and a transaction id. The support of each k-sequence is determined by the temporal join of k-1 sequences that share a common suffix. Both Apriori and GSP make multiple database scans. GSP also uses a hash structure. The innovation of the SPADE algorithm is that it decomposes the original problem into smaller problems and uses lattice search techniques to solve the problems independently in memory. The important contribution of this algorithm is that all sequences are discovered in only three database scans. Zaki et al. propose an algorithm that finds frequent sequences that always preceded Plan failure with applications in emergency situations. The algorithm (Plan Mine) is based on the SPADE algorithm and it successively filters out uninteresting sequences. In the first step, sequences that are not related to plan failures are filtered out. In the second step, it filters out frequent sequences that appear frequently in plan failures, but also have high support in plans that did not fail. In the third step, redundant patterns are removed, which are patterns that have sub patterns in both good and bad plans. In the final step, dominated patterns are removed. These are patterns that have subsequence that have lower support in good plans and higher support in bad plans than the original sequence in which they are contained.

- **The SPAM and I-SPAM Algorithms:** The authors [8] propose a sequential pattern mining technique that utilizes a bitmap representation called SPAM. The algorithm is the first sequential mining method that utilizes a depth-first approach to explore the search space. Combining this search strategy with an effective pruning technique that reduces the number of candidates makes the algorithm particularly suitable for very long sequential patterns. However, the algorithm needs that the whole database can be stored in main memory, which is the main outcome of the algorithm. As sequences are generated visiting each node in the tree, two sorts of children are generated from every node: sequence-extended sequences (sequence extension step or S-step) and itemset-extended sequences (item-extension step or I-step). Finally, an efficient representation of the data is used, which is a vertical bitmap representation. The bitmap is created for each item as follows: If item A belongs to transaction j, then a 1 is recorded for A in transaction j. Otherwise a 0 is recorded. SPAM was compared against SPADE and PrefixSpan using several small, medium, and large data sets. On small datasets, SPAM was about 2.5 times faster than SPADE, but PrefixSpan was a bit faster than SPAM on very small data sets. On large data sets, SPAM was faster than PrefixSpan by about an order of magnitude. The authors attributed the fast performance on large data sets to its efficient bitmap-based representation. An improved version of SPAM is proposed is called I-SPAM.In this version, approximately half of the maximum memory requirement of SPAM is needed with no sacrifice in the performance time.

- **Generalized Sequential Patterns (GSP):** In 1996 GSP Algorithm was introduce by Shrikant and Agarawal for Sequential Access Pattern Mining. It uses the downwards-closure property of sequential patterns and adopts a multiple-pass, candidate generate and-test approach. This algorithm is outlined as follows: In the first scan of database, it finds all of the frequent items > = minimum support. Each such item yields a 1-event frequent sequence consisting of that item. Each subsequent pass starts with seed set of sequential patterns the set of sequential pattern found in previous pass. This seed is used to generate new potentially frequent patterns, called candidate sequences. Each candidate sequence contains one more item then the seed sequence pattern from which it was generated. The number of instances of items in a sequence is the length of the sequence. So, all of the candidate sequences in the given pass will have the same length. The algorithm terminates when no unique sequence pattern is found in a iteration, or no new candidate sequence number can be generated. But there is some problem of minimum support count with this algorithm as compared to PrefixSpan Algorithm.
- **PrefixSpan Algorithm:** Prefix Span comes under pattern growth method for mining sequential patterns [8]. Its major idea is that, instead of projecting sequence databases by considering all the possible occurrences of frequent subsequence's, the projection is based only on frequent prefixes because any frequent subsequence can always be found by growing a frequent prefix. This way, sequential patterns are grown in each projected database by exploring only local frequent sequences. Three major steps of prefix span are as follows:
  1. Find all 1-itemset sequential patterns by scanning the database WASD.
  2. Divide the search space to get projected databases: The furnished set of serial access patterns can be partitioned into the following three subsets according to the three prefixes: i) the ones with prefix ***a***, ii) the ones with prefix ***b*** iii) the ones with prefix ***c.***
  3. Find subset of sequential patterns: The subsets of sequential patterns can be mined by constructing the corresponding set of projected databases and mining each recursively. Experimental Execution time of GSP and PrefixSpan Algorithm by varying Minimum support as in Table 1 :

Table I.

Minimum Support Threshold (Frequency of web pages)	**GSP Algorithm Execution Time (in ms)**	**Prefix Span Algorithm Execution Time (in ms)**
3	7313	218
6	2359	125
9	1047	94
12	562	87
15	438	78
18	328	31
21	234	15

III. COMPARATIVE ANALYSIS OF ALGORITHM PERFORMANCE

[8] The depicted Table shows the comparative analysis of all algorithms.

Table II. Tabular Taxonomy of Sequential Pattern mining Algorithm

	Apriori-Based		Pattern Growth						Early-Pruing
Algorithm	**Generate & Test**	**Multiple Scan of Database**	**Sampling and/or compression**	**Candidate Sequence pruning**	**Search Space partitioning**	**Depth First Traversal**	**Prefix Growth**	**Memory Only**	**Vertical Position of DB**
Apriori	x	x	X		X				
GSP	x	x		X					
SPADE	x			X	X	X			X
FreeSpan					X			x	
PrefixSpan				X	X		x	x	
SPAM	x					X		x	x
I-SPAM	x					X		x	

IV. CONCLUSION

Recent research has mostly focused on Web usage analysis partly because of its applicability in e-business. We expect privacy issues, distributed Web mining, and Semantic Web mining to attract equal, if not more, interest from the research community Increased use of Web usage mining Techniques will require that privacy issues be addressed however. Similarly, aggregating data in a central site and then mining it is rarely scalable..Finally, researchers will need to leverage the semantic information. Exposing content semantics web and the link explicitly can help in many tasks, including mining the hidden Web—that is, data stored in databases and not accessible through search engines.In this work, we study the possible use of the pattern clicking capabilities to classify the web traffic. The findings of truthful knowledge, user data and server log patterns allows Web based organizations to mining user access patterns and helps in future developments, planning and also to target more famous advertising campaigns aimed at grouped users. We can conclude that; to identify common patterns in Web the clicking pattern algorithms need to develop in large scale distributive environments for giving furnished knowledge (patterns) from web usage data. The proposed work can be done in high density distributive environment for analysing essential knowledge from the massive web log.

REFERENCES

[1] Etminani, K., Delui, A.R., Yanehsari, N.R. and Rouhani, M., "Web Usage Mining: Discovery of the Users' Navigational Patterns Using SOM", First International Conference on Networked Digital Technologies, Pp.224-249, 2009.

[2] Nina, S.P., Rahman, M., Bhuiyan, K.I. and Ahmed, K., "Pattern Discovery of Web Usage Mining", International Conference on Computer Technology and Development, Vol. 1, Pp.499-503, 2009.

[3] Chih-Hung Wu, Yen-Liang Wu, Yuan-Ming Chang and Ming-Hung Hung, "Web Usage Mining on the Sequences of Clicking Patterns in a Grid Computing Environment", International Conference on Machine Learning and Cybernetics (ICMLC), Vol. 6, Pp. 2909-2914, 2010.

[4] Maratea, A. and Petrosino, A., "An Heuristic Approach to Page Recommendation in Web Usage Mining", Ninth International Conference on Intelligent Systems Design and Applications, Pp. 1043-1048, 2009.

[5] Nasraoui, O., Soliman, M., Saka, E., Badia, A. and Germain, R., "A Web Usage Mining Framework for Mining Evolving User Profiles in Dynamic Web Sites", IEEE Transactions on Knowledge and Data Engineering, Vol. 20, No. 2, Pp. 202-215, 2008.

[6] Wu, K.L., Yu, P. S. and Ballman, A., "SpeedTracer: A Web usage mining and analysis tool", IBM Systems Journal, Vol. 37, No. 1, Pp. 89-105, 1998. A Survey on Web Usage Mining Global Journal of Computer Science and Technology Volume XI Issue IV Version 1 ©2011 Global Journals Inc. (US)

[7] Tzekou, P., Stamou, S., Kozanidis, L. And Zotos, N., "Effective Site Customization Based on Web Semantics and Usage Mining", Third International IEEE Conference on Signal-Image Technologies and Internet-Based System, Pp.51-59, 2007.

[8] Manan Parikh, Bharat Chaturvedi, and chetna chand ,"A Study of Sequential Pattern Mining Algorithms ", International Journal of Application of Innovation in Engineering and Management, Vol 2, Pp. 103-109, 2013.

[9] Zaki, M., "SPADE: An Efficient Algorithm for Mining Frequent Sequences", Machine Learning, vol. 42, no .1/2, pp.31–60, 2001.

