Volume 5, Issue 9, September 2015 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 Existing Web Log Mining Algorithms

Paridhi Nigam*, Pallavi Jain
SVITS, RGPV,
M.P., India

Abstract--This paper presents a review of existing web log mining algorithms. Nowadays web log mining is a very popular and computationally expensive tasks. Generally the web mining term is related to the extraction of the useful patterns from the web data. The data mining tools and techniques are used for mining web log data. We have also explained the fundamentals of web log mining using the concept of frequent item set mining.

Keywords— Web mining, web log mining, patterns, techniques, tools

I. INTRODUCTION

The web mining is used to extract the useful information from the World Wide Web by using data mining techniques. The overall tasks under web mining are generally divided into three main categories. These are.

I. Web content mining
II. Web structure mining
III. Web usage mining

The first one is the web content mining is used to search the web pages by using the content of the web pages as search words. The second web structure mining is the collection of methods which are used for mining or extracting the structure or hierarchy or the links of a website.The web usage mining is related to the application of data mining tools and techniques on the web to discover the web user patterns. It helps organizations in increasing the user satisfaction. The web usage mining consists of three major steps. These steps are as follows:

I. Preprocessing
II. Pattern discovery
III. Pattern analysis

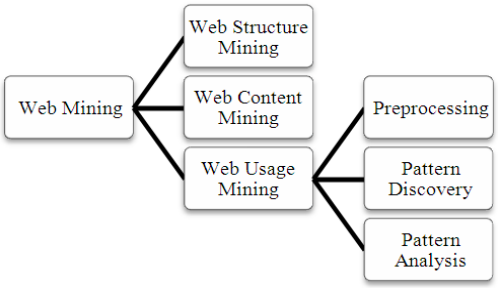

Fig. 1 Web mining categorization

In preprocessing step of web usage mining, the data is extracted from the web data set & then this data is preprocessed. In preprocessing, the noise is removed from the data. The output of preprocessing phase contains information like, how many pages accessed, which is accessed how many times, which user accessed which page, access time, access date, access duration etc.

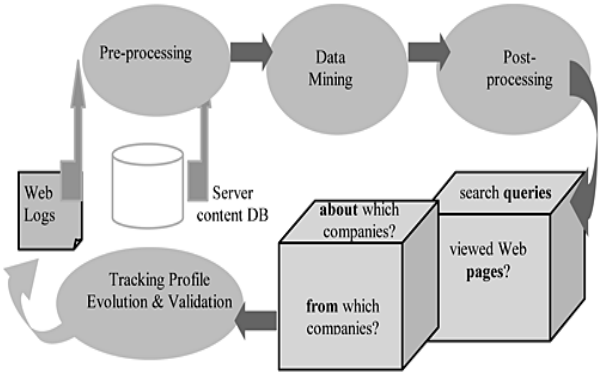

Fig. 2 Web usage mining process

II. LITERATURE SURVEY

The authors Alexandra's Nanopoulos et al [2] proposed a web mining method based on the concept of web perfecting. It helped in the decreasing the user latency perception ratio.

Mathis Gerry et al [6] proposed three different web mining approaches. These three methods are based on association rules, frequent sequences, and frequent generalized sequences. The authors have developed and implemented the algorithms for all three methods. Association rule learning [10] is a very common learning method from discovery of useful patterns from data & also for representing the useful patterns in form of a rule.

Both Charm and Closet [8,9] inherit the same data structures and computing framework of their big brothers declat and FP-Growth respectively. They implement Algorithm 4, but they differ in the way the closed frequent itemsets are stored in order to exploit the Sub-sumption Lemma. Charm adopts a hash table, were the hash function is the sum of the transactions ids supporting an itemset. Closet uses a trie-like structure, indexed by a two-level hash. The first level is based on the last item of the itemset to be checked and the second on its support. FP-Close [12] is inspired to Closet, thus using the same divide et impera approach and same FP-tree data structure. What makes FP-Close different from other CFIM algorithms is the application of the projecting approach to the historical collection of closed frequent itemsets. Not only a small dataset is associated to each node of the tree, but also a pruned subset of the closed itemsets mined so far is forged and used for duplicate detection. Indeed, this technique is called progressive focusing and it was introduced by [10] for mining maximal frequent itemsets. Together with other optimizations, this truly provides dramatic speed-up, making FP-Close order of magnitudes faster than Charm and Closet, and also making it worth to be celebrated as the fastest algorithm at the FIMI workshop 2003 [11].

Chi et al. [13] propose an algorithm called *Moment* for mining frequent closed itemsets over data streams. It uses a *CET Tree (Closed Enumerate Tree)* to maintain the main information of itemsets. Each node in CET Tree represents an itemset with different node type. Some nodes in CET Tree are not closed so that there are still some redundant nodes in CET Tree. *Moment* must maintain huge CET nodes for a frequent closed itemset. Chi et al. indicated that the ratio of CET nodes for a closed itemsets is about 20:1. If there are a large number of frequent closed itemsets, it will consume a lot of memory space. When a new transaction arrives, the node is inserted and updated according to its node type. The exploration of frequent itemsets and node type checking are time consuming. *CFI-Stream* is another algorithm for this problem [14]. Only the closed itemsets are maintained in a lexicographical ordered tree which is called *DIU Tree (Direct Update Tree)*. Each node consists of a closed itemset and its support count. When a new transaction X arrives, CFI-Stream will generate all the subsets of X, and check if each subset Y is closed or not after the transaction arrives. To check whether an itemset Y is closed or not, CFI-Stream may need to search all supersets of Y from *DIU Tree*. It takes a lot of time to generate all the subsets of a new transaction and search their supersets from DIU Tree.

III. CONCLUSION

Web usage mining is concerned with the extracting the users pattern from the web log data set. Generally the data mining techniques like association rule, frequent item mining on the web log to gain the information related to the web users search patterns. In this paper, the review of web usage mining is proposed.

REFERENCES

[1] Alexandros Nanopoulos, Dimitris Katsaros and Yannis Manolopoulos "*Effective prediction of web-user accesses: A data mining approach*," in Proc. Of the Workshop WEBKDD, 2001.

[2] Mathias Gery, Hatem Haddad "*Evaluation of Web Usage Mining Approaches for Users Next Request Prediction*" WIDM‟03 Proceedings of the 5th ACM international workshop on web information and data management p.74-81, November 7-8, 2003.

[3] Piatetsky-Shapiro, G. (1991), Discovery, analysis, and presentation of strong rules, in G. Piatetsky-Shapiro & W. J. Frawley, eds, „Knowledge Discovery in Databases‟, AAAI/MIT Press, Cambridge.

[4] J. Pei, J. Han, and R. Mao. Closet: An efficient algorithm for mining frequent closed itemsets. In DMKD '00: ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery, pages 21–30, May 2000.

[5] M. J. Zaki and C.-J. Hsiao. Charm: An efficient algorithm for closed itemset mining. In SDM '02: Proceedings of the second SIAM International Conference on Data Mining, April 2002.

[6] G. Grahne and J. Zhu. Efficiently using prefix-trees in mining frequent itemsets. In FIMI '03: Proceedings of the ICDM 2003 Workshop on Frequent Itemset Mining Implementations, November 2003.

[7] G. Grahne and J. Zhu. Fast algorithms for frequent itemset mining using fptrees. IEEE Transactions on Knowledge and Data Engineering, 17(10):1347– 1362, 2005.

[8] Chi, Y., Wang, H., Yu, P.S., Muntz, R.R.: Moment: Maintaining Closed Frequent Itemsets over a Stream Sliding Window. In: Proceedings of 2004 IEEE International Conference on Data Mining, Brighton, pp. 59–66 (2004).

[9] Jiang, N., Gruenwald, L.: CFI-Stream: Mining Closed Frequent Itemsets in Data Streams. In: Proceedings of 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Philadelphia, PA, USA, pp. 592–597 (2006).

