



International Journal of Emerging Technologies in Computational and Applied Sciences (IJETCAS)

www.iasir.net

Review on Path Traversal for Web Navigation Mining

Sagar More¹, Suraj Patil², Nihit Agrawal³, Rohit Pachlor⁴

^{1,2,3,4}Department of Computer Engineering
SVKM's NMIMS University, Mumbai

Mukesh Patel School of Technology, Management and Engineering, Shirpur Campus, Maharashtra,
INDIA

Abstract: Web Navigation Pattern is point comes under Web Usage Mining which shows how one can visited from one page to another i.e. it shows navigational behaviour. Mostly this pattern mining is success part of e-commerce and mobile commerce. Analysing this data will help the organizations to realize the lifetime value of their clients, and provide them with a more sophisticated structure of the web site and services. So Web Navigation Pattern can be easily shows the trends towards web pages that visited next for browsing session. This paper will show Web Usage Mining, Web Navigation Patterns generation, Frequent patterns generated, Web logs and throughout surfing patterns. Here we proceed through introduction of web concepts and their navigation patterns, next related work point covers different techniques for navigation mining and pattern generation. Next path traversal graph covers the concept of links and via-links and shows how the visitor follows the path to visit different web pages. Web Navigation Pattern Generation point covers some different pattern mining techniques

Keywords: Web Navigation Mining, Traversal Pattern, Frequent Paths, Browsing Sessions, Path Traversal Graph

I. Introduction

World Wide Web or Web is the largest and popular source that is easily available, reachable and accessible at low cost, provides quick response to the users and reduces the burden on the users of physical movements. However, there is too much information on the Internet and too many activities to monitor. For analyzing such huge information for intelligent system it may create problem regarding performance. With vast amount information being shared worldwide, there was a requisite to find required information in a systematic and effective way.

While website visitor visited web pages from one web page to another web page, certainly it creates a path. That path shows how the user traverses from one web page to another web page directly. There is also one condition is that visitors wish to one web page but there may be no direct web page access so visitor has to follow via-links which shows indirect access of web page. Hence for whole scenario of web accessing Web Navigation Pattern is basic factor to identify intention of user. A series of Web pages in a website requested by a visitor in a single visit is referred to as a session.

For pattern identification structure of website is very rarely considered. Support count for each web pages changes according to their location from home page, i.e. web page which located near to home page have higher support count than the web page that located away. Main thing is that single traversal pattern cannot show whole intention of user navigation behavior. A user sequence is a complete traversal sequence for a user from entering to exiting the web system. Web Servers gathered information in the form of web logs. For mining patterns from web logs there are different methods available in the form of algorithms. Practically that mined patterns are useful for improvement of web site design, improvement of system performance for building of adaptive websites. Web Navigation Pattern mining from web logs is mostly based on association rule, concept of maximal forward references and support count of web page.

In this paper, we go through web mining approach to discover the patterns to be formed by web site's visitor which is known as throughout-surfing patterns. Maximal forward reference concept is utilized to discover the paths. At the first stage a compact structure called the path traversal graph is constructed to trace the tracks of Web navigation. Then we can apply graph traversal algorithm to discover the throughout surfing patterns. In our mining approach there is no generation of candidate pattern and also it scans the database only once to maintain efficiency. Hence efficiency of proposed mining approach is better than the traditional sequential approach.

II. Related Work

To find the intentions of website visitor's and according to that intentions manage the website and its webpages is important in e-commerce. So that web navigation pattern mining is basic operation to improve its efficiency. There are different techniques available to find web navigation pattern. But main thing is first it generate frequent traversal pattern. Here it uses the concept of maximal forward references. Jian Pei, J. Han, B. Mortazavi and Hua Zhu [7] found that Web Log Mining helps in extracting interesting and useful patterns from the Log File of the sever. Mainly three parameters are considered to discover Web access patterns: association rules, sequential patterns, and clustering [8]. Linear genetic programming is a variant of the GP technique that acts on linear genomes. MAFTP (Maintenance of Frequent traversal patterns) is the technique to maintain frequent traversal patterns when inserted in database about the positions of candidate traversal sequences so that it reduces the number of scans [1]. Knowledge discovery technique for web usage data for this research is mainly focused on three basic parameters: association rules, sequential patterns, and clustering [10]. Association rules are the form $X \Rightarrow Y$ where the rule body X and rule head Y are set of items within set of transactions. The rule $X \Rightarrow Y$ states that the transactions which contain the items in X are likely to contain in Y also. Sequential patterns are used to discover frequent subsequences in Web Usage Mining to find sequential navigation patterns that appear in users_ sessions frequently. And clustering is technique that groups data based on distance function where similarity between groups is main parameter. One technique known as position coded technique to construct Pre-Order Linked WAP-tree [11]. Like FP tree WAP stores original web access sequences that firstly stores web access database on a prefix tree. Proposed algorithm for this frequent header node links of original WAP tree in a pre-order fashion and find the ancestor or descendant and relationships between nodes of the tree. Common pattern mining is sequential pattern mining to discover all maximal frequent sequential patterns.

III. Path Traversal Graph

By using some topology, extraction of user's navigation pattern for server logs is comes under web usage mining. Web usage mining has materialized as fundamental procedure for comprehending more custom-made, user approachable and business oriented intelligent top web services. And this patterns find are stored in the form of graph termed as a path traversal graph, to record information about the navigation paths of website visitors [5]. Actually website is the collection of connected hyperlinks and by using directed graph, a website is represented where webpages are connected with edges corresponding to documents and hyperlinks respectively. In the application of mining throughout-surfing patterns, as we predict where a visitor will go at any node, we need to know first where the visitor comes from. Accordingly, the concept of via-links is introduced in this paper to record the "from-to-via" information in the proposed graph, which is unique to the mining of throughout-surfing patterns. Therefore, we propose a novel data structure called path traversal graph consisting of a set of vertices, edges, and via-links to store the information from Web browsing sessions. The compact structure of the path traversal graph can help improve the efficiency of mining throughout-surfing patterns.

IV. Web Navigation Pattern Generation

Stochastic Model for Sequential Data:

This model is based on the assumption that the value of the element in a sequence is subject to the value of the previous elements [13]. Standardly Markov chain is defined a set of state $S = \{s_1, s_2, \dots, s_k\}$ that drive a step-by-step process in which we move from one state to the other at given time instants $t = 1, 2, \dots, n$ according to a $k \times k$ transition matrix P so that P_{ij} represents the probability of moving from state i to state j .

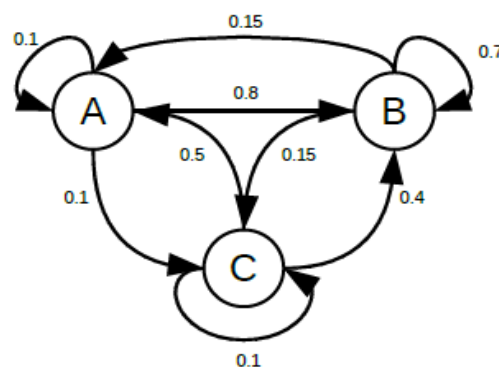


Figure 1: A Markov chain with three states and their corresponding transition probabilities.

Also there is optional Hidden Markov Model where current state is not known i.e. is in hidden form but it can predict from observed predict sequences.

Incremental Web Traversal Pattern Mining:

This technique mine web traversal patterns incrementally and here previous pattern is considered for new pattern to be discovered hence benefit is that mining time is reduced. So here focus on well storage structure which shows previous mining result is important. Lattice structure is selected to keep the previous mining results. The lattice structure is a well storage structure. It can quickly find the relationships between patterns [3].

Utility based pattern mining:

After path traversal pattern mining, a concept called utility comes out, to mining model where, utility is a show how item set is useful for pattern recognition. By introducing the concept of utility into web path traversal pattern mining, end user's preference is subjective value, and the objective value could be the browsing time a user spent by user on a webpage. So main focus of this pattern mining is finds the Web traversal sequences that have high utility beyond a minimum threshold [6].

Apriori algorithm for Pattern Mining:

The Apriori algorithm to find frequent trajectory pattern mining is based on the sequential pattern mining framework. Trajectory is considered to be tracking the paths of an object. A trajectory can be represented as a set of coordinates in sequence in Euclidian space [9].

V. Conclusion

This paper provides an overview of web navigation pattern generation. From generated graph we introduced the concept of path traversal graph to record information about navigation path of website visitor's. Throughout surfing pattern is the term superset of web traversal pattern which is the effective to display trends towards next visited web pages. Also paper focused on pattern searching techniques which in turn important for discovering the different patterns and intention to find interest of web user. Other result shows that throughout-surfing patterns is effective for content management and are applicable to providing surfing paths and personalized the configuration of web site.

VI. Future Scope

In future we can focus on effectiveness and efficiency to improve the performance of system. So finally we reach throughout surfing pattern which shows high performance for the system. Another interesting issue is generation of path traversal graph for incremental database.

References

- [1] Show-Jane Yen, Yue-Shi Lee, Chung-Wen Cho, "An Efficient Approach for the Maintenance of Path Traversal Patterns", Proceedings of the 2004 IEEE International Conference on e-Technology, e-Commerce and e-Service (EEE'04).
- [2] Jatin Chhugani, Nadathur Satish, Changkyu Kim, Jason Sewall, and Pradeep Dubey, "Fast and Efficient Graph Traversal Algorithm for CPUs: Maximizing Single-Node Efficiency", 2012 IEEE 26th International Parallel and Distributed Processing Symposium.
- [3] Show-Jane Yen, Yue-Shi Lee and Min-Chi Hsieh, "An Efficient Incremental Algorithm for Mining Web Traversal Patterns", Proceedings of the 2005 IEEE International Conference on e-Business Engineering (ICEBE'05).
- [4] Zhixiang Chen, Richard H. Fowler, Ada Wai-Chee Fu, Chunyue Wang, "Linear and Sublinear Time Algorithms for Mining Frequent Traversal Path Patterns From Very Large Web Logs", Proceedings of the Seventh International Database Engineering and Applications Symposium (IDEAS'03).
- [5] Yao-Te Wang, Anthony J.T. Lee b, "Mining Web navigation patterns with a path traversal graph", Expert Systems with Applications 38 (2011) 7112–7122
- [6] Lin Zhou, Ying Liu, Jing Wang, Yong Shi, "Utility-based Web Path Traversal Pattern Mining", Seventh IEEE International Conference on Data Mining – Workshops.
- [7] Jian Pei, Jiawei Han, Behzad Mortazavi-asl, Hua Zhu, "Mining Access Patterns Efficiently from Web Logs".
- [8] Ajith Abraham, Vitorino Ramos, "Web Usage Mining Using Artificial Ant Colony Clustering and Genetic Programming".
- [9] Arthur.A.Shaw, N.P. Gopalan, "Frequent Pattern Mining of Trajectory Coordinates using Apriori Algorithm", International Journal of Computer Applications (0975 – 8887) Volume 22– No.9, May 2011.
- [10] Federico Michele Facca, Pier Luca Lanzi, "Mining interesting knowledge from weblogs: a survey", Data & Knowledge Engineering 53 (2005) 225–24.
- [11] C.I. EZEIFE, YI LU, "Mining Web Log Sequential Patterns with Position Coded Pre-Order Linked WAP-Tree", Data Mining and Knowledge Discovery, 10, 5–38, 2005 Springer Science + Business Media, Inc.
- [12] Rakesh Agrawal, Ramakrishnan Srikant, "Mining Sequential Patterns".
- [13] Anna Gutowska, Luis L. Perez, "A Comparison of Methods for Classification and Prediction of Web Access Patterns", COMP540 - Final Report - Spring 2010.

Acknowledgments

I take this as an opportunity to extend my sincere thanks to Prof. Suraj Patil for offering me unique platform to learn exposure and knowledge in field of web navigation mining. I am also thankful to Prof. Nihit Agrawal, Prof.Rohit Pachlor who helps me in information gathering in field of data mining. Also, I would like to thanks all the authors whose papers are referred by me in completing this review. Their efforts put up in the papers, helped me to understand the subject up to a great extent.