WEB USAGE MINING USING GSP ALGORITHM: A STUDY ON SULTANAH BAHIYAH LIBRARY ONLINE DATABASES

Name: Yousef Abd- ALMohdi Hazzaimeh

Universiti Utara Malaysia

2008

WEB USAGE MINING USING GSP ALGORITHM: A STUDY ON SULTANAH BAHIYAH LIBRARY ONLINE DATABASES

A Thesis is submitted to college Arts & Sciences in partial

fulfillment of the requirement for the degree master

(Intelligent system)

University Utara Malaysia

By

Name: Yousef Abd- ALMohdi Hazzaimeh

Metric No:89300

KOLEJ SASTERA DAN SAINS
(College of Arts and Sciences)
Universiti Utara Malaysia

PERAKUAN KERJA KERTAS PROJEK
(Certificate of Project Paper)

Saya, yang bertandatangan, memperakukan bahawa
(I, the undersigned, certify that)

YOUSEF ABD-AL MOHDI HAZZAIMEH
(89300)

calon untuk Ijazah
(candidate for the degree of) **MSc. (Intelligent System)**

telah mengemukakan kertas projek yang bertajuk
(has presented his/her project paper of the following title)

WEB USAGE MINING USING GSP ALGORITHM:
A STUDY ON SULTANAH BAHIYAH LIBRARY ONLINE DATABASES

seperti yang tercatat di muka surat tajuk dan kulit kertas projek
(as it appears on the title page and front cover of project paper)

bahawa kertas projek tersebut boleh diterima dari segi bentuk serta kandungan dan meliputi bidang ilmu dengan memuaskan.
(that the project paper acceptable in form and content, and that a satisfactory knowledge of the field is covered by the project paper).

Nama Penyelia Utama
(Name of Main Supervisor): **ASSOC. PROF. DR NORITA MD. NORWAWI**

Tandatangan
(Signature) : ____________________

Tarikh
(Date) : 24/6/08

PERMISSION TO USE

In presenting this thesis in partial fulfilment of the requirements for a postgraduate degree from Universiti Utara Malaysia, I agree that the University Library may make it freely available for inspection. I further agree that permission for copying of this thesis in any manner, in whole or in part, for scholarly purpose may be granted by my supervisor or, in their absence by the Dean of the Graduate Studies. It is understood that any copying or publication or use of this thesis or parts thereof for financial gain shall not be allowed without my written permission. It is also understood that due recognition shall be given to me and to Universiti Utara Malaysia for any scholarly use which may be made of any material from my thesis.

Requests for permission to copy or to make other use of materials in this thesis, in whole or in part, should be addressed to

Dean of Graduate Studies
Universiti Utara Malaysia
06010 UUM Sintok
Kedah Darul Aman.

ABSTRACT

Application of data mining to the World Wide Web referred as Web mining is at the cross road of research from several research communities which can be divided into three branches: Web Content Mining, Web Structure Mining and Web Usage Mining. Sultanah Bahiyah Library which is considered as one of the most important resources for University Utara Malaysia (UUM) students provides several online databases that can be utilized by its user's in seeking the needed information. Analyzing the usage or access pattern of these databases is time consuming and is not an easy task because the number of users accessing the site every day are too many. The goals of this study are to propose a suitable technique for preprocessing web log data of Sultanah Bahiyah Library online databases that can reduce the file size and to analyze the user's access pattern of the online databases using web usage mining. In this study web usage mining use sequential pattern technique with GSP algorithm. This study found out that Emeraldinsight was visited most by 20% of the user. And the top three sequences were {Emeraldinsight, Epnet, Proquest_direct} with support = 16.6%.

ABSTRAK

Penggunaan perlombongan data ke atas laman Web dipanggil perlombongan Web yang kini berada di persimpangan jalan pernyelidikan dari pelbagai komuniti penyelidik terbahagi kepada tiga cabang: Perlombongan Kandungan Web, Perlombongan Struktur Web dan Perlombongan Maklumat Penggunaan Web. Perpustakaan Sultanah Bahiyah merupakan sumber penting bagi para pelajar Universiti Utara Malaysia menyediakan beberapa pangkalan data atas talian yang boleh digunapakai oleh pengguna untuk mencari maklumat yang diperlukan. Analisa maklumat penggunaan atau corak capaian pangkalan data tersebut memerlukan masa yang lama dan bukanlah satu tugas yang mudah memandangkan capaian hariannya begitu banyak. Matlamat kajian ini adalah untuk mencadangkan teknik prapemprosesan data log web pangkalan data atas talian Perpustakaan Sultanah Bahiyah yang berupaya mengecilkan saiz fail data dan menganalisa corak capaian pengguna menggunakan perlombongan maklumat penggunaan web. Dalam kajian ini, kaedah perlombongan maklumat penggunaan web adalah menggunakan teknik "Corak Perlombongan Tersusun" dengan algoritma GSP. Pangkalan data Emerald Insight didapati paling banyak dilawat oleh hampir 20% pengguna . Manakala turutan bagi tiga pangkalan data yang teratas adalah Emeral Insight, Epnet dan Proquest Direct dengan nilai sokongan 16%.

Acknowledgment

قال العماد الأصبهاني :" إني رأيت أنه لا يكتب إنسان كتاباً في يومه إلاّ قال في غده : لو غُيّر هذا لكان أحسن ، ولو
زيد كذا لكان يُستحسن ، ولو قدّم هذا لكان أفضل ، ولو تُرك هذا لكان أجمل ، وذلك من أعظم العبر ، وهو دليل استيلاء
النقص على جملة البشر " .

"All praises and thanks to ALLAH"

I would like to thank my supervisor ASSOC. PROF. DR NORITA MD. NORWAWI who encouraged and helped me very much in doing this research.

great thanks form my heart to my parents for their prayers and my brothers, sisters, friends, and colleagues for their encouragements and support.

Finally, I would like to thank all lecturers, administration, and a stuff of Universiti Utara Malaysia for their help and support.

TABLE OF CONTENTS

CHAPTER TWO: LITERATURE OVERVIEW

CHAPTER THREE: METHODOLOGY

CHAPTER FOUR: FINDINGS AND DISSCUSSIONS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

Acronym	Meaning
DISC	Direct Sequence Comparison
DM	Data mining
GSP	Generalized Sequential Pattern
KDD	Knowledge Discovery in Database
LOGML	Log Markup Language
MDR	Mining Data Records
SPADE	Sequential Pattern Discovery using Equivalent Class
WAMF	Web Access Monitoring and Filtering
WWW	World Wide Web
XGMML	Extensible Graph Markup and Modeling Language

CHAPTER 1

INTRODUCTION

1.1 Introduction

Data mining techniques are widely used for retrieving the related and hidden information and at the same time to enhance the way that these databases work by looking for more suitable and comfortable environments for its user's. Sultanah Bahiyah Library which is considered as one of the most important resources for University Utara Malaysia (UUM) students provides several online databases that can be utilized by its user's in seeking the needed information. As known, access record to these online databases can be obtained from the server's web log that contains a lot of data possibly needed by user. A web log is a listing of page reference data (Dunham, 2002). However it may contain unnecessary information. This unnecessary information can be minimized or reduced by using web usage mining through mining process of the web log. Web usage mining can be used for many different purposes by looking at the sequence of pages of user access in order to evaluate and update the log structure.

5.3 Future Work

Lastly, for future work, other methods for analyzing sparse data can be used in the study of Web log access, use different similarity sequential pattern technique, and explore other different techniques or algorithm on the same problem using the same data.

References

Adriaans, P., & Zantinge, D. (1997). *Data mining*: Addison-Wesley Longman Publishing Co., Inc. Boston, MA, USA.

Agrawal, R., & Srikant, R. (1995). *Mining sequential patterns* .Paper presented at the Proceedings of the Eleventh International Conference on Data Engineering.

Albanese, M., Picariello, A., Sansone, C., & Sansone, L. (2004). *A web personalization system based on web usage mining techniques* .Paper presented at the Proceedings of the 13th international World Wide Web conference on Alternate track papers & posters.

Antunes, C., & Oliveira, A. (2004). *Sequential pattern mining algorithms: Trade-offs between speed and memory* .Paper presented at the Proceedings of the Second Workshop on Mining Graphs, Trees and Sequences at the 15th European ECML and the 8th European PKDD.

Chiu, D., Wu, Y., & Chen, A.(2004). An Efficient Algorithm for·Mining Frequent Sequences by a New Strategy without Support Counting .*Contact,* 3,15.

Cooley, R., Mobasher, B., & Srivastava, J. (1997). *Web mining: information and pattern discovery on the World Wide Web* .Paper presented at the Tools with Artificial Intelligence, 1997. Proceedings., Ninth IEEE International Conference on.

Do, T. Chang, K. and Hui, S. C. (2004). Web mining for cyber monitoring and filtering. Cybernetics and Intelligent Systems, 2004 IEEE Conference on IEEE, vol.1, pp. 399 – 404.

Dunham, M. (2002). Data Mining: Introductory and Advanced Topics: Prentice Hall PTR Upper Saddle River, NJ, USA.

Eirinaki, M., & Vazirgiannis, M. (2003). Web mining for web personalization. ACM Transactions on Internet Technology (TOIT), 3(1), 1-27.

Eljilani, E.M. (2007). Web usage pattern extraction. Unpublished MSC intelligent system. Dissertation University Utara Malaysia.

El-Sayed, M., Ruiz, C., & Rundensteiner, E. (2004). *FS-Miner: efficient and incremental mining of frequent sequence patterns in web logs* .Paper presented at the Proceedings of the 6th annual ACM international workshop on Web information and data management.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). The KDD process for extracting useful knowledge from volumes of data .*Communications of the ACM,* Han, J., Pei, J., & Yan, X. (2005). Sequential Pattern Mining by Pattern-Growth: Principles and Extensions .*STUDIES IN FUZZINESS AND SOFT COMPUTING,* 180-183.

Hsu, J. (2002). *WEB MINING: A Survey of World Wide Web Data Mining Research and Applications* .Paper presented at the Decision Sciences Institute 2002 Annual Meeting Proceedings.

Huang, X., Cercone, N., & An, A. (2002). *Comparison of interestingness functions for learning web usage patterns* .Paper presented at the Proceedings of the 11th International Conference on Information and Knowledge Management (CIKM'02).

Kosala, R., & Blockeel, H. (2000). Web mining research: a survey. *ACM SIGKDD Explorations Newsletter, 2*(1), 1-15.

Khasawneh, N., & Chan, C. (2005 .(*Web usage mining using rough sets* .Paper presented at the Fuzzy Information Processing Society, 2005. NAFIPS 2005. Annual Meeting of the North American.

Krishnaswamy, S. Loke, S. W. and Zaslavsky, A. (2002). Web and e-business application: Application run time estimation: a quality of service metric for web-based data mining services. Proceedings of the 2002 ACM symposium on Applied computing SAC '02. ACM pp. 1153 – 1159.

Leleu, M., Rigotti, C., Boulicaut, J., & Euvrard, G. (2003). GO-SPADE: Mining Sequential Patterns over Datasets with Consecutive Repetitions .*LECTURE NOTES IN COMPUTER SCIENCE* 293-306.

Li, H., Zhang, D., Hu, J., Zeng, H., & Chen, Z. (2007). *Finding keyword from online broadcasting content for targeted advertising* .Paper presented at the Proceedings of the 1st international workshop on Data mining and audience intelligence for advertising.

Liu, B., Grossman, R., & Zhai, Y. (2003). *Mining data records in Web pages* .Paper presented at the Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining.

Masseglia, F., Poncelet, P., & Teisseire, M. (1999). Using data mining techniques on Web access logs to dynamically improve hypertext structure .*ACM SIGWEB Newsletter,* 8(3), 13-19.

Dunham, M. (2002). *Data Mining: Introductory and Advanced Topics* :Prentice Hall PTR Upper Saddle River, NJ, USA.

Mobasher, B., Jain, N., Han, E., & Srivastava, J. (1996). Web mining: Pattern discovery from world wide web transactions *.Dept. Comput. Sci., Univ. Minnesota, Minneapolis, MN, Tech. Rep. TR-96-050.*

Pei, J., Han, J., Mortazavi-Asl, B., Wang, J., Pinto, H., Chen, Q., et al. (2004). Mining Sequential Patterns by Pattern-Growth: The PrefixSpan Approach *.IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING.*1440-1424 ,

Plantevit, M., Laurent, A., & Teisseire, M. (2006). *HYPE: mining hierarchical sequential patterns* .Paper presented at the Proceedings of the 9th ACM international workshop on Data warehousing and OLAP.

Punin, J., Krishnamoorthy, M., & Zaki, M. (2003). We*b Usage Mining-Languages and Algorithms* .Paper presented at the Exploratory Data Analysis in Empirical Research: Proceedings of the 25th Annual Conference of the Gesellschaft Für Klassifikation EV, University of Munich, March 14-16, 2001.

Ren, J., & Zhou, X. (2006). A New Incremental Updating Algorithm for Mining Sequential Patterns *.Journal of Computer Science,* 2(4), 318-321.

Srikant, R., & Agrawal, R. (1996). Mining Sequential Patterns: Generalizations and Performance Improvements *.LECTURE NOTES IN COMPUTER SCIENCE* 3(17).

Srivastava, J., Cooley, R., Deshpande, M., & Tan, P. (2000). Web usage mining: discovery and applications of usage patterns from Web data *.ACM SIGKDD Explorations Newsletter,* 1(2), 12-23.

Stumme, G., Hotho, A., & Berendt, B. (2002) *.Usage mining for and on the semantic web* .Paper presented at the National Science Foundation Workshop on Next Generation Data Mining.

Zaïane, O. (1999). Principles of Knowledge Discovery in Databases, CMPUT690. University of Alberta, Department of Computing Science.

Zaki, M. (2001). SPADE: An Efficient Algorithm for Mining Frequent Sequences . *Machine Learning*, 42(1), 31-60.

Zheng, Q., Xu, K., Ma, S., & Lv, W. (2002). The Algorithms of Updating Sequential Patterns *,Arxiv preprint cs.DB/0203027.*