WEB USAGE MINING FOR UUM LEARNING CARE USING ASSOCIATION RULES

A project submitted to the Graduate School in partial fulfillment of the requirements for the degree Master of Science (Intelligent System)
Universiti Utara Malaysia

By:
Azizul Azhar bin Ramli

© Azizul Azhar bin Ramli, 2004
All rights reserved.

JABATAN HAL EHWAL AKADEMIK
(*Department of Academic Affairs*)
Universiti Utara Malaysia

PERAKUAN KERJA KERTAS PROJEK
(Certificate of Project Paper)

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

AZIZUL AZHAR

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

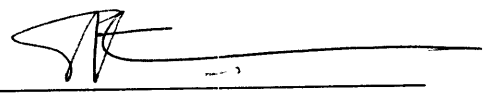
telah mengemukakan kertas projek yang bertajuk
(has presented his/ her project paper of the following title)

WEB USAGE MINING FOR UUM LEARNING CARE USING ASSOCATION RULES

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 filed is covered by the project paper).

Nama Penyelia Utama
(Name of Main Supervisor): **MR. MOHD. SHAMRIE SAININ**

Tandatangan
(Signature) : 

Tarikh
(Date) : 26/06/04

PERMISSION TO USE

In presenting this thesis in partial fulfillment of the requirement 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 purposes may be granted by my supervisor or, in her absence, by the dean of the Graduate School. It is also understood that due recognition shall be given to me and Universiti Utara Malaysia for any scholarly use which may be made of any material from my thesis.

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

Dean of Graduate School,
Universiti Utara Malaysia
06010 UUM Sintok
Kedah Darul Aman
Malaysia

ABSTRAK

Ledakan maklumat di dalam gerbang Web menjadikannya pilihan terbaik bagi penyelidikan perlombongan data. Aplikasi bagi teknik perlombongan data bagi gerbang Web dirujuk sebagai perlombongan Web dan telah digunakan dalam tiga pendekatan yang berbeza; Perlombongan Isi Web, Perlombongan Struktur Web dan Perlombongan Penggunaan Web. Pembelajaran Elektronik adalah salah satu aplikasi di dalam gerbang Web dimana ianya akan berhadapan dengan jumlah data yang sangat besar. Bagi menghasilkan corak penggunaan portal dan amalan pengguna, kajian ini akan mengimplimentasikan proses aras tinggi bagi perlombongan penggunaan Web menggunakan algoritma asas bagi Peraturan Kesatuan – algoritma *Apriori*. Perlombongan penggunaan Web mengandungi tiga fasa utama iaitu Preprosesan Data, Penjelajahan Corak dan Analisis Corak. Sumber utama iaitu data mentah adalah terdiri daripada fail-fail log pelayan dan ianya perlu melalui fasa-fasa dalam perlombongan penggunaan Web bagi menghasilkan keputusan akhir – set peraturan. Dengan keupayaan teknik perlombongan data, pendekatan perlombongan penggunaan Web telah digabungkan dengan asas Peraturan Kesatuan, algoritma *Apriori* bagi mengoptimumkan kandungan portal Pembelajaran Elektronik universiti. Akhir sekali, kajian ini akan membentangkan keputusan serta analisisnya supaya pihak pengurusan Web boleh menggunakan segala keputusan tersebut untuk tindakan bernilai yang sewajarnya.

ABSRACT

The enormous content of information on the World Wide Web makes it obvious candidate for data mining research. Application of data mining techniques to the World Wide Web referred as Web mining where this term has been used in three distinct ways; Web Content Mining, Web Structure Mining and Web Usage Mining. E-Learning is one of the Web based application where it will facing with large amount of data. In order to produce the university E-Learning (UUM Educare) portal usage patterns and user behaviors, this paper implements the high level process of Web usage mining using basic Association Rules algorithm – Apriori Algorithm. Web usage mining consists of three main phases, namely Data Preprocessing, Pattern Discovering and Pattern Analysis. Main resources, server log files become a set of raw data where it's must go through with all the Web usage mining phases to produce the final results – set of rules. With the powerful of data mining technique, Web usage mining approach has been combined with the basic Association Rules, Apriori Algorithm to optimize the content of the university E-Learning portal. Finally, this paper will present an overview of results with the analysis and Web administrator can use the findings for the suitable valuable actions.

ACKNOWLEDGEMENTS

First of all, I would like to express my appreciation to Allah, the Most Merciful whom granted me the ability and willing to start and complete this project. I pray to his greatness to inspire and to enable me to continue the work for benefits of my religion, Islam and country.

I would also like to express my gratitude to my supervisor, Encik. Mohd. Shamrie bin Sainin, lecturer of the Artificial Intelligent Department of Information Technology Faculty at the Universiti Utara Malaysia for her excellent guidance and advice through completing this project. Many special thanks to the Information System Officer from UUM Computer Centre, Cik Roslina Hanafiah who honestly prepared and provided the required data for this project, others lecturers with their informal guidance and friends for their helps and supports. Especially for my beloved family and my dearest one whose encourage me much, throughout this project. Thanks for every never-endings support and kindness. May Allah bless us, InsyaAllah.

Only God knows everything!
Thank you.

AZIZUL AZHAR BIN RAMLI
Artificial Intelligent Department,
Information Technology Faculty, Universiti Utara Malaysia

TABLE OF CONTENTS

CHAPTER	DESCRIPTIONS	PAGE
	PERMISSION TO USE	**i**
	ABSTRAK	**ii**
	ABSTRACT	**iii**
	ACKNOWLEDGEMENT	**iv**
	TABLE OF CONTENTS	**v**
	LIST OF FIGURES	**viii**
	LIST OF TABLES	**ix**
1	**PROJECT INTRODUCTION**	**1**
	1.1 Project Background	1
	1.2 Problem Statement	4
	1.3 Project Objectives	5
	1.4 Project Significant	6
	1.5 Project Scope and Boundaries	6
	1.6 Thesis Organization	7
2	**LITERATURE REVIEW**	**8**
	2.1 Data Mining	8
	2.2 Web Mining	10
	2.3 Web Usage Mining	12
	2.4 High Level Web Usage Mining Process	14

2.4.1 Data Preprocessing..... 14
2.4.2 Pattern Discovery..... 18
2.4.3 Pattern Analysis..... 20
2.5 Web/Server Log Files..... 22
2.5.1 Web/Server Log Files Analysis..... 23
2.6 Association Rules..... 25
2.6.1 *Apriori* Algorithm..... 26
2.6.2 *Apriori* Algorithm Process Flow..... 28
2.6.3 *Apriori* Algorithm Property..... 29
2.7 Web Usage Mining in E-Learning Field..... 29

3 SERVER LOG FILES..... 31
3.1 Server Log Files Overview..... 31
3.2 Web Server Log..... 32
3.3 A Web Server Log Data Primer..... 34
3.4 Demographic Server Log Data..... 35
3.5 Performance Server Log Data..... 35
3.6 Example of Server Log Files..... 36
3.7 Server Log Files Analysis..... 38

4 PROJECT METHODOLOGY AND IMPLEMENTION..... 40
4.1 Project Methodology..... 40
4.2 Project Implementation..... 43
4.2.1 Server Log Files..... 43
4.2.2 Data Selection..... 44
4.2.3 Data Preprocessing..... 45
4.2.4 Pattern Discovery - Association Rules (*Apriori* Algorithm)..... 49
4.2.5 Pattern Analysis..... 51
4.2.6 Results..... 54

5 **FINDINGS AND RESULTS**..... 55

5.1 General Pattern Analysis Results (*access pattern and users behaviors – descriptive statistic*)..... 55

5.2 Association Rules Results (*supports and confidences of the different level – Apriori algorithm*)..... 58

VI **CONCLUSION AND RECOMMENDATION**..... 65

REFERENCES..... 67

APPENDIXES..... 68

A: Source Code for Rearrange Log Data by *Clent_IP* 68

B: Sample of Server Log Files for UUM Educare Portal..... 71

C: Sample of Clean Server Log Files for */educare* Path..... 77

D: Sample of Clean Server Log Files for */educare/portfolio* Path with the Provided Options..... 83

E: Sample of Clean Server Log Files for */educare/portfolio/dms* option Path..... 89

F: Sample of Clean Server Log Files for */educare/portfolio/dms/dmsget.php* Path..... 95

LIST OF FIGURES

NO	FIGURES	DESCRIPTIONS	PAGE
1.	Figure 1.1	Taxonomy of Web Mining.....	3
2.	Figure 2.1	A High Level Web Usage Mining Process.....	14
3.	Figure 2.2	Sample Web Server Log Files.....	22
4.	Figure 2.3	Basic *Apriori* Algorithm.....	27
5.	Figure 2.4	*Apriori* Algorithm Process Flow.....	28
6.	Figure 4.1	Suggested Project Methodology.....	41
7.	Figure 4.2	Sample of Raw Server Log Files.....	43
8.	Figure 4.3	Main *ARunner* 1.0 User Interface.....	50
9	Figure 4.4	*Apriori* Output for *ARunner* 1.0 User Interface.....	51
10.	Figure 4.5	Main *WebLog Expert 3.0* User Interface.....	53
11.	Figure 4.6	Main *Sawmill 6.5.4* User Interface.....	53
12.	Figure 5.1	Most Requested Options on UUM Educare Portal.....	56
13.	Figure 5.2	UUM Educare Portal for Daily Countries Activity.....	57
14.	Figure 5.3	Output for UUM Educare Options Association Rules (*related options*).....	60
15.	Figure 5.4	Six Most Accepted Rules for UUM Educare (*related options*).....	61
16.	Figure 5.5	Output for UUM Educare Options Association Hyperedges (*orderly archived*).....	62
17.	Figure 5.6	Six Most Accepted Rules UUM Educare Options Association Hyperedges (*orderly archived*).....	63

LIST OF TABLES

NO	TABLES	DESCRIPTIONS	PAGE
1.	Table 2.1	Sample Web Server Log Files (*after preprocessing process*).....	24
2.	Table 3.1	Typical Navigation and Activity Server Log Data.....	34
3.	Table 4.1	Suggested Web Usage Mining Phases with the Sub Tasks.....	41
4.	Table 4.2	Preprocessed Server Log Files.....	46
5.	Table 4.3	Sub Tasks for Server Log Files Data Preprocessing Phase.....	47
6.	Table 5.1	Support and Confidence for *~educare/portfolio* with UUM Educare Options.....	59
7.	Table 5.2	Support and Confidence for *~educare/portfolio/dms* Option Path.....	64

CHAPTER 1

INTRODUCTION

This chapter will discuss about the project background that contains the general overview of the project includes the brief description about the data mining technologies and Web usage mining as a part of Web mining approach. In addition, the project background sub chapters also describe the tools and software that was used for this project and also the approach that is used for the analysis purposes. The description of the project problem statement, lists of the project objectives and details of project scope and boundaries are also discuses in this chapter. Finally, the thesis organization that contains the structure of chapters that is included in this report.

1.1 Project Background

Finding useful patterns in data is known by different names (including data mining) in different communities (e.g., knowledge extraction, information discovery, information harvesting, data archeology and data pattern processing) (Fayyad *et al.*, 1996). Data mining is a burgeoning and promising research field in computer science field. Data mining is a technique used to deduce useful and relevant information to guide professional decisions and other scientific research (Chen, Han and Yu, 1996). The objective of data mining is to identify valid novel, potentially useful and understandable correlations and patterns in existing data (Chung and Gray, 1999). It is a cost-effective

The contents of the thesis is for internal user only

REFERENCES

Abd. Wahab, M. H, Siraj, F and Yusoff, N. (2004). *Log Mining Using Generalize Association Rules*. In Proceedings of Master Final Project 2004 Presentation, UUM, Malaysia.

Agrawal, S, Agrawal, R., Deshpande, P.M. Gupta, A. Naughton, J., Ramakrishna, R and Sarawagi, S. (1996). *On the Computation of Multidimensional Aggregates*. In Proc. of the 22nd VLDB Conference, Mumbai, India. Pp 506-521.

Agrawal, R., Imielinski, T. and Swami, A. (1993). *Mining Association Rules between Sets of Items in Large Databases*. In Proceedings of the International ACM SIGMOD Conference, Washington DC, USA, pages 207–216.

Agrawal, R. and Srikant, R. (1994). *Fast Algorithm for Mining Association Rules*. Proc. of the 20th VLDB Conference. Pp 487-499.

Agrawal, R., and Srikant, R. (1995). *Mining Sequential Patterns*. In Proc. of the Eleventh International Conference on Data Engineering (ICDE), Taiwan. Pp 3-14.

Bertot, J. C., McClure, C. R., Moen, W. E., and Rubin, J. (1997). *Web Usage Statistics: Measurement Issues and Analytical Techniques*. Government Information Quarterly. 14 (4). Pp 373-395.

Borgelt, C. (2004). *Apriori: Finding Association Rules/Hyperedges with the Apriori Algorithm*. School of Computer Science, University of Magdeburg.

Boon Lay, C, Khalid, M and Yusof, R. (!999). *Intelligent Database by Neural Network and Data Mining*. In Proc. of Artificial Intelligent Applications in Industry, Kuala Lumpur. Pp 201-219.

Buchner, A. G., and Mulvenna, M. D. (1998). *Discovering Internet Marketing Intelligence Through Online Analytical Web Usage Mining*. SIGMOD Record, 27 (4), Pp 54-61.

Chen, M.-S., Jan, J., Yu, P.S. (1996). *Data Mining: An Overview from a Database Perspective*. IEEE Transactions on Knowledge and Data Engineering, (8:6). Pp 866.883.

Chung, H. M., Gray, P. (1999). *Special Section: Data Mining*. Journal of Management Information Systems, (16:1). Pp 11-17.

Cooley R. (2000). *Web Usage Mining: Discovery and Application of Interesting Patterns from Web Data*. PhD thesis, Dept. of Computer Science, University of Minnesota.

Cooley R., Mobasher B., and Srivastava J. (1997). *Web Mining: Information and Pattern Discovery on the World Wide Web*. In Proceedings of the 9th IEEE International Conference on Tools with Artificial Intelligence (ICTAI'97).

Cooley, R., Mobasher, B. and Srivastava, J. (1997). *Grouping Web Page References into Transactions for Mining World Wide Web Browsing Patterns*. Technical Report TR 97-021, University of Minnesota, Dept. of Computer Science, Minneapolis.

Cooley, R, Mobasher, B. and Srivastava, J. (1999). *Data Preparation for Mining World Wide Web Browsing Patterns*. Knowledge and Information Systems, 1(1).

Dereson, C. (1997). *Using an Incomplete Data Cube as a Summary Data Sieve*. Bulletin of the IEEE Technical Committee on Data Engineering. Pp 19-26.

Edelstein, H., A. (2001). *Pan for Gold in the Clickstream*. Informationweek.com, March12. 2001, Pp 77-91.

Fayyad, U., Piatetsky-Shapiro, G., and Smyth, R. (1996). *The KDD Process for Extracting Useful Knowledge from Volumes of Data*. Communications of the ACM, (39:11). Pp 27-34.

Gray, J., Bosworth, A., Layman, A and Pirahesh, H. (1996). Data Cube: A Rational Aggregation Operator Generalizing Group-By, Cross-Tab and Sub-Totals. In IEEE 12th International Conference on Data Engineering. Pp 152-159.

Han, J., Kamber, M. (2001). *Data Mining: Concepts and Techniques*. Morgan-Kaufmann Academic Press, San Francisco.

Hand, D. J. (1998). *Data Mining: Statistics and More*". The American Statistician. May (52:2). Pp 112-118.

Harinarayan, V, Rajaraman, A. and Ullman J.D. (1996). *Implementing Data Cubes Efficiently*. In Proc. of 1996 ACM-SIGMOD Int. Conf. Management of Data. Montreal, Canada. Pp 311-322.

Jiang, Q. (2003). *Web Usage Mining: Process and Application*. Presentation for CSE 8331.

Kosala, R., Blockeel, H. (2000). *Web Mining Research: A Survey*. ACM SIGKDD (Special Interest Group on Knowledge Discovery and Data Mining) Explorations. June, (2:1). Pp 1-10.

Mannila, H., Toivonen, H. and Verkamo, A. I. (1994). *Efficient Algorithms for Discovering Association Rules*. In AAAIWorkshop on Knowledge and Discovery in Databases, Seattle, Washington, USA, Pp 181–192.

McLaughlin, M., Goldberg, S. B., Ellison, N., and Lucas, J. (1999). *Measuring Internet Audiences: Patrons of an On-line Art Museum*. In S. Jones (Ed.), Doing Internet Research: Critical Issues and Methods for Examining the Net. Thousand Oaks, CA: Sage. Pp. 163-178.

Peacock, P. R. (1998). *Data Mining in Marketing: Part 1*. Marketing Management, Winter. Pp 9-18.

Pitkow, J., and Bharat, K. (1994). *Webvis: A Tool for World Wide Web Access Log Analysis*. In First International WWW Conference.

Rajagopalan, B., Krovi, R. (2002). *Benchmarking Data Mining Algorithms*. Journal of Database Management, Jan-Mar. 13, Pp 25-36.

Brin, S., Motwani, R., Ullman, J. D. and Tsur, S. (1997). *Dynamic Itemset Counting and Implication Rules for Market Basket Data*. In Proceedings of the International ACM SIGMOD Conference, Tucson, Arizona, USA, Pp 255–264.

Shukla, A., Deshpande, P.M., Naughton, J and Ramaswamy, K. (1996). *Storage Estimation for Multidimensional Aggregates in the Presence of Hierarchies*. In Proc. of the 22nd VLDB Conference. Mumbai, India. Pp 522-531.

Spiliopoulou M. (1999). *Data mining for the Web*. In Proceedings of Principles of Data Mining and Knowledge Discovery, Third European Conference, PKDD'99, Pp588-589.

Srivasta, J., Cooley, R., Deshpande, M., and Tan P. N. (2000). *Web Usage Mining: Discovery and Application of Web Usage Pattern from Web Data*. Department of Computer Science and Engineering, University of Minnesota.

Tang, C.; Lau, R.W.H.; Li, Q.; Yin, H.; Li, T.; and Kilis, D.(2000). *Personalized Courseware Construction Based on Web Data Mining*. In Proc. of the First International Conference on Web Information Systems Engineering (WISE 2000) vol.2, Pp. 204-211.

Tang, Y. T. and McCalla, G. (2001). *Student modeling for a Web based Learning Environment: a Data Mining Approach*. Department of Computer Science, University of Saskatchewan, Canada.

Webopedia. (2001). *Web Server*. http://webopedia.internet.com/TERM/W/Web_server.html Date Accessed : 06 April 2004.

Wilson, T. (1999). *Web Trafic Analysis Turns Management Data to Business Data. TechWeb*. http://www.internetk.com/story/INW19990402S0006 Date Accessed : 24 March 2004.

Xue, G. R., Zeng, H. J., Ma, W. Y and Lu, C. J. (2002). *Log Mining to Improve the Performance of the Methods from statistic, Neural Nets, Machine Learning and Experts System*. Morgan Kaufman.

Zaiane, O. and Luo, J. (2001). *Towards Evaluating Learners' Behavior in a Web-based Distance Learning Environment*. In *Proc*. of IEEE International Conference on Advanced Learning Technologies, Madison, WI. Pp 357-360.