LOG MINING USING GENERALIZED ASSOCIATION RULES

A project submitted to the Faculty of Information Technology in partial fulfillment of the requirements for the degree
Master of Science (Intelligent Knowledge Based System)
Universiti Utara Malaysia

by
Mohd Helmy Abd Wahab

PERMISSION TO USE

In presenting this project 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 project in any manner, in whole or in part, for scholarly purposes may be granted by my supervisor(s) or, in their absence, by the Dean of the Graduate School. It is understood that any copying or publication or use of this theses 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 project.

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

Dean of Faculty on Information Technology
Universiti Utara Malaysia
06010 UUM Sintok
Kedah Darul Aman

ABSTRACT (ENGLISH)

Explosive growth in size and usage of the World Wide Web has made it necessary for Web site administrators to track and analyze the navigation patterns of Web site visitors. To achieve this goal, the use of web mining tool is necessary. Web mining can be defined as the use of data mining techniques to automatically discover and extract information from web documents. Since Data Mining is primarily concerned with the discovery of knowledge and aims to provide answers to questions that people do not know how to ask, it is not an automatic process. Rather one has to exhaustively explores very large volumes of data to determine otherwise hidden relationships. The process extracts high quality information that can be used to draw conclusions based on relationships or patterns within the data. However, data mining techniques are not easily applicable to Web data due to problems both related with the technology underlying the Web and the lack of standards in the design and implementation of Web pages. Information collected by the Web servers are kept in the server log is the main source of data for analyzing user navigation patterns. Once logs have been pre-processed and sessions have been obtained, there are several kinds of access pattern mining that can be performed depending on the needs of the analyst. Since the method use in this study relied on relatively simple techniques therefore the information gathered is adequate for real user profile data due to the noise in the data has to be first tackled. In this study, Data Mining techniques known as generalized association rules was used in order to get some insights into website usage pattern. For the purpose of this study, server logs from tutor.com portal were retrieved, pre-processed and analyzed. An important finding from this study is that Mathematics subject generally popular from UPSR, PMR and UPSR levels. On the contrary, arts subjects are not popular to Tutor.com users. The system administrator may consider evaluating the content and the link for such subjects, so that the real problem can be identified.

ABSTRACT (BAHASA MELAYU)

Perkembangan yang pesat di dalam penggunaan *World Wide Web* (www) menyebabkan perlunya pentadbir laman web menjejaki dan menganalisis corak pelayaran pengunjung laman web. Untuk mencapai matlamat ini, penggunaan alat bantu pengguna web amat diperlukan oleh pentadbir laman web. Perlombongan web didefinisikan sebagai penggunaan teknik perlombongan data untuk memenuhi dan memperolehi maklumat daripada dokumen web. Perlombongan Data adalah tertumpu kepada penemuan pengetahuan dan juga bertujuan untuk menyediakan jawapan kepada soalan pengguna, teknik ini tidak dapat melakukan pemprosesan secara automatik. Sebaliknya, pelayan web perlu meneroka data yang banyak untuk menentukan perkaitan antara data. Proses ini mengekstrak maklumat berkualiti yang boleh digunakan untuk membuat sesuatu kesimpulan berdasarkan hubungan atau paten yang terdapat didalam data berkenaan. Walaubagaimanapun, teknik perlombongan data tidak mudah diaplikasikan kepada data web disebabkan masalah yang berkaitan dengan teknologi web dan kekurangan keseragaman dalam rekabentuk dan implementasi laman web. Maklumat yang terkumpul oleh pelayan web akan disimpan didalam log pelayan yang mana merupakan sumber utama data bagi menganalisis corak pelayaran pengguna. Setelah log pelayan di pra-proses dan sesi pengguna diperolehi, perlombongan paten capaian boleh dilaksanakan dengan beberapa kaedah. Ini bergantung kepada matlamat penganalisis log pelayan. Kaedah yang digunakan dalam kajian ini adalah berdasarkan kepada teknik yang lazim digunakan, maka maklumat yang dikumpulkan tidak mencukupi untuk menjadikan data profil pengguna kerana "*noise*" yang terdapat dalam data perlu di pra-proses terlebih dahulu. Teknik perlombongan data yang dikenali sebagai *generalized association rules* digunakan untuk mendapatkan paten penggunaan laman web. Untuk kajian ini, log pelayan dari portal Pendidikan Utusan telah diperolehi, diproses dan telah dianalisis menggunakan alat bantu perlombongan data. Dapatan kajian menunjukkan, bahawa Matematik merupakan subjek yang popular untuk peringkat UPSR, PMR dan peringkat UPSR.. Sebaliknya, subjek sastera tidak begitu popular kepada pengguna Portal Pendidikan. Oleh sebab itu, penyelenggara Portal Pendidikan Utusan perlu mempertimbangkan kandungan serta pautan kepada setiap subjek di dalam Portal Pendidikan supaya masalah yang sebenarnya dapat diatasi.

ACKNOWLEDGEMENT

Heartfelt thanks are due first to my main supervisor, Associate Professor Fadzilah Siraj for patiently navigating and generously sharing her rich source of knowledge with me. She is indeed a teacher of "*open hand, open mind, and open heart*".

Equally thankful to my second supervisor, Miss Nooraini Yusoff for zealously giving a hand to her utmost. I also sincerely acknowledge the cooperation and consideration given by Mr. Hassani Hassan from Bahagian Pendidikan Utusan Melayu (Malaysia) Sdn. Bhd.

I am most indebted to my parents for all the love and support in giving me the best gift; "*Education*"- a lifelong priceless present that can never be destroyed by calamities. Last but not least, let me express my deep appreciation to all who lend a hand in materializing this project.

Mohd Helmy Abd Wahab
Faculty of Information Technology
2004

TABLE OF CONTENTS

CHAPTER 3: WEB SERVER LOG

CHAPTER 4: METHODOLOGY

CHAPTER 5: IMPLEMENTATION

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

INTRODUCTION

1.1 Context of the study

With the explosive growth of data available on the World Wide Web (WWW), discovery and analysis of useful information from the World Wide Web becomes a practical necessity. Data Mining is primarily concerned with the discovery of knowledge and aims to provide answers to questions that people do not know how to ask. It is not an automatic process but one that exhaustively explores very large volumes of data to determine otherwise hidden relationships. The process extracts high quality information that can be used to draw conclusions based on relationships or patterns within the data.

Using the techniques used in Data Mining, Web Mining applies the techniques to the Internet by analyzing server logs and other personalized data collected from customers to provide meaningful information and knowledge. Web access pattern, which is the sequence of accesses pursued by users frequently, is a kind of interesting and useful knowledge in practice (Pei, 2000). Today web browsers provide easy access to myriad sources of text and multimedia data. More than 1 000 000 000 pages are indexed by search engines, and finding the desired information is not an easy task (Pal *et al.*, 2002). Web Mining is now a popular term of techniques to analyze the data from World Wide Web (Pramudiono, 2004). A widely accepted definition of the web mining is the application of data mining techniques to web data. With regard to the type of web data, web mining can be classified into three types: Web Content Mining, Web Structure Mining and Web Usage Mining.

The contents of the thesis is for internal user only

REFERENCES

Agrawal, R. and Srikant, R. (1994). Fast algorithms for mining association rules. *Proc. of the 20th VLDB Conference*. pp 487 - 499.

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

Bestavros, A. (1995). Using Speculation to Reduce Server Load and Service Time On The WWW. In *Proceedings of the fourth ACM International Conference on Information and Knowledge Management*. pp. 403 – 410.

Borgelt, C. and Kruse, R. (2002). Induction of Association Rules: Apriori Implementation. *15th Conference on Computational Statistics (CompStat 2002)*

Borges, J. and Levene, M. (1999). Data Mining of User Navigation Patterns. *Proceedings of the WEBKDD '99 Workshop on Web Usage Analysis and User Profiling*. pp. 31 – 36.

Borges, M. (2000). A Data Mining Model to Capture User Web Navigation Patterns. *PhD Thesis*. University of London.

Cadez, I., Heckerman, D., Meek, C., Smyth, P., and White, S. (2000). Visualization of Navigation Patterns On a Web Site Using Model Based Clustering. In *Proceedings of the sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.*

Chakrabarti, S., Dom, B., Gibson, D., Klienberg, J., Kumar, S., Raghavan, P., Rajagopalan, S., and Tomkins, A. (1999). Mining the Link Structure of The World Wide Web. *IEEE Computer*. Vol. 32. No. 8. pp. 60 – 67.

Chen, M. S., and Park, J. S., and Yu, P. S. (1996). Data Mining for Path Traversal Patterns in A Web Environment. *16th International Conference on Distributed Computing Systems*. pp. 385 – 392.

Cheeseman, P. and Stutz, J. (1996). Bayesian classification (autoclass): Theory and results. In U.M. Fayyad, G. Piatetsky-Shapiro, P. Smith, and R. Uthurusamy, editors, *Advances in Knowledge Discovery and Data Mining*. AAAI/MIT Press. Pp. 153 -180.

Cooley, R., Mobasher, B., and Srivastava, J. (1997).Web Mining: Information and Pattern Discovery on the World Wide Web. *Technical Report TR 97-027.*

Cooley, R., Tan, P. –N., and Srivastava, J. (1999). Discovery of Interesting Usage Patterns from Web Data. *Technical Report TR 99-022.*

Davidson, B. D. (2001). Web Traffic Logs: An Imperfect Resource for Evaluation. *Ninth Annual Conference of The Internet Society.*

Desikan, P., Srivastava, J., Kumar, V., Tan, P. N. (2002). Hyperlink Analysis – Techniques & Applications. *Army High Performance Computing Center Technical Report.*

Drott, M. C. (1998). Using Web Server Logs to Improve Site Design. *Association for Computing Machinery (ACM) Proceeding of the Sixteenth Annual International Conference on Computer Documentation. pp. 43 – 50.*

Dunham, M. H. (2002). *Data Mining: Introductory and Advanced Analysis*. New Jersey: Prenhall.

Fayyad, U. M., Piatetsky-Shapiro, G., Smyth, P., and Uthurusamy, R. (1996). Advances

in Knowledge Discovery and Data Mining. *AAAI Press/ The MIT Press.*

Fisher, D. (1995). Optimization and simplification of hierarchical clusterings. *Proc. of the First Int'l Conference on Knowledge Discovery and Data Mining.* Pp. 118-123.

Haigh, S. and Megarity, J. (1998). *Measuring Web Site Usage: Log File Analysis.* Network Notes #57.

Han, J., Cai, Y., and Cercone, N. (1993). Data-driven discovery of quantitative rules in relational databases. *IEEE Transactions on Knowledge and Data Eng.* Vol. 5. pp. 29-40.

Kerkhofs, J., Vanhoof, K., and Pannemas, D. (2001). Web Usage Mining on Proxy Server: A Case Study. *Technical Report. Limburg University Centre.*

Kaufman, L. and Rousseeuw, P. J. (1990). *Finding Groups in Data: an Introduction to Cluster Analysis.* John Wiley & Sons.

Kosala, R. and Blockeel, H. (2000). Web Mining Research: A Survey. *ACM SIGKDD.* Vol. 2. Issue 1.pp. 1 – 14.

Lee, R. S. T. and Liu, J. N. K. (2001). iJADE eMiner: A Web-Based Mining Agent Based on Intelligent Java Agent Development Environment (iJADE) on Internet Shopping. *PAKDD 2001.* LNAI 2035. pp. 28 – 41.

Lin, W. Alvarez, S. A., and Ruiz, C. (2001). Efficient Adaptive – Support Association Rule Mining for Recommender Systems. *Kluwer Academic Publisher.*

Madria, S., Bhowmick, S. S., Ng, W. K., and Lim, E. P. (1999). Research Issue in Web Data Mining. *Data Warehousing and Knowledge Discovery.*

Mannila, H., Toivonen, H., and Verkamo, A. I. (1995). Discovering frequent episodes in Sequences. *Proc. Of the First Int'l Conference on Knowledge Discovery and Data Mining.* pp. 210 – 215.

Mehta, M., Agrawal, R., and Rissanen, J. (1996). SLIQ: A fast scalable classifier for data mining. *Proc. of the Fifth Int'l Conference on Extending Database Technology.*

Moen and McClure. (1997). An Evaluation of the U.S. GILS Implementation. URL: http://www-lan.unt.edu/slis/research/gilseval.htm Date Accessed 30 January 2004.

Mobasher, B., Jain, E., Han, E., and Srivastava, J. (1996). Web mining: Pattern discovery from World Wide Web Transactions. *Technical Report TR 96-050.*

Mobasher, B., Cooley, R., and Srivastava, J. (1999). Creating adaptive web sites through usage-based clustering of URLs. *In Proceeding of the 1999 IEEE Knowledge and Data Engineering Exchage Workshop (KDEX'99) (Nov.).*

Moh, C-H., Lim, E-P., Ng, W. K. (2000). DTD-Miner: A Tool for Mining DTD from XML Documents. *WECWIS 2000.* pp. 144 – 151.

Mohammadian, M. (2001). Intelligent Data Mining and Information Retrieval from World Wide Web for E-Business Applications. http://www.ssgrr.it/en/ssgrr2002w/papers/230.pdf

Nakayama, T., Kato, H., and Yamane, Y. (2000). Discovering the Gaps Between Web Site Designers' Expectations and Users' Behaviour. *Proc. Of the Ninth Int'l World Wide Web Conference.*

Nasraoui, O., Frigui, H., Joshi, A., and Krishnapuram, R. (1999). Mining Web Access Logs uing a Fzzy Rlational Custering Agorithm Bsed on ARobust Estimator. In *Proceedings of the eighth International World Wide Web Conference.*

Nasraoui, O.and Petenes, C. and (2003). An Intelligent Web Recommendation Engine Based on Fuzzy Approximate Reasoning.

Ng, R. and Han, J. (1994). Efficient and effective clustering method for spatial data mining. *Proc. of the 20th VLDB Conference*. pp. 144-155.

Novak and Hoffman. (1996). *New Metrics for New Media: Toward the Development of Web Measurement Standards.* http://www2000.ogsm.vanderbilt.edu/novak/web.standards/webstand.html [Date Accessed: 28 February 2004].

Padmanabhan, V. N. and Mogul, J. C. (1996). Using Predictive Prefetching to Improve World Wide Web Latency. *ACM SIGCOMM Computer Communications Review*, 26(3). pp. 22 – 36.

Pal, S. K., Talwar, V., and Mitra, P. (2002). Web Mining in Soft Computing Framework: Relevance, State of the Art and Future Directions. *IEEE Transactions on Neural Networks.*

Pramudiono, I. (2004). Parallel Platform for Large Scale Web Usage Mining. *Phd Thesis*. Department of Computer Science, University of Tokyo.

Pei, J., Han, J., Asl, B. M., and Zhu, H. (2000). Mining Access Patterns Efficiently from Web Logs.

Perkowitz, M. and Etzioni, O. (1998). Adaptive sites: Automatically Synthesizing Web Pages. *Proceedings of the fifteenth National Conference on Artificial Intelligence*. pp. 727 – 732.

Perkowitz, M. and Etzioni, O. (2000). Towards Adaptive Web Sites: Conceptual

Framework and Case Study. *Artificial Intelligence*. Vol. 118. pp.245 – 275.

Perotti, V. (2003). Techniques for Visualizing Website Usage Patterns With an Adaptive Neural Network. *The ACM Digital Library*. Pp 35 – 40.

Ravid, G., Yaffe, E., and Tal, E. (2002). Web Mining in Education: Using Students' Log Files as an Indicator of On-Line Learning and as a Tool for Improving On-Line Instruction. http://www.infosoc.haifa.ac.il/kennes/Gilad3.doc. [Date Accessed: 20 March 2004].

Rosenfeld, L. and Morville, P. (1998). Information Architecture for the World Wide Web. O'Reilly, Cambridge.

Sarukkai, R. R. (2000). Link prediction and path analysis using Markov chains. In *Proceedings of the ninth International World Wide Web Conference.*

Shahabi, C., Zarkesh, A. M., Adibi, J., and Shah, V. (1997). Knowledge Discovery from Users Web-page navigation. *Workshop on Research Issue in Data Engineering.*

Spiliopoulou, M. and Faulstich, L. C. (1998). WUM: A Web Utilization Miner. *EDBT Workshop WebDB98*

Srikant, R., Vu, Q., and Agrawal, R. (1997). Mining Association Rules with Item Constraints. *American Association of Artificial Intelligence (AAAI).*

Srivastava, J., Desikan, P., and Kumar, V. (2002). Web Mining: Accomplishments and Future Directions.

Srivastava, J., Cooley, R., Tan, P. –N. (2000). Web Usage Mining: Discovery and

Applications of Usage Patterns from Web Data. *SIGKDD Explorations*. Vol. 1. No. 2. pp. 12 – 33.

Stout, R. (1997). *Web Site Stats: tracking hits and analyzing traffic*. Osborne McGraw-Hill: Berkeley.

Tao, F., Murtagh, F., and Farid, M. (2003). Weighted Association Rule Mining using Weighted Support and Significant Framework. *SIGKDD 2003.*

Toolan, F., and Kuhmerick, N. (2002). Mining Web Logs for Personalized Site Maps. *First International Wokshop on Mining for Enhanced Web Search.*

Weiss, S. M. and Kulikowski, C. A. (1991). Computer Systems that Learn: Classification and Prediction Methods from Statistics, Neural Nets, Machine Learning, and Expert Systems. Morgan Kaufmann.

Wang and Liu, H. (1998). Discovering Typical Structures of Documents: A Roadmap Approach. *Proceeding of the ACM SIGIR Symposium on Information Retrieval.*

Weiss, S. M. and Kulikowski, C. A. (1991). Computer Systems that Learn: Classification and Prediction Methods from Statistics, Neural Nets, Machine Learning, and Expert Systems. Morgan Kaufmann.

Wilson, T. (1999). Web Traffic Analysis Turns Management Data to Business Data. *TechWeb*. http://www.internetk.com/story/INW19990402S0006 Date Accessed: 25 February 2004

Wong, C., Shiu, S., and Pal, S. (2001). Mining Fuzzy Association Rules for Web Access Case Adaptation. *Proceeding of the Workshop Program at The Fourth International Conference on Case Based Reasoning 2001.*

Wu, K. –L., Yu, P. S., and Ballman, A. (1998). SpeedTracer: A Web Usage Mining and Analysis Tool. *IBM System Journal.* Vol. 37. No. 1.

Xue, G. R., Zeng, H. J., Chen, Z., Ma, W. Y., and Lu, C. J. (2002). Log Mining to Improve the performance of Site Search. *Third Int. Conf. of WISEw '02.*

Yang, Q. (2002). Building Association Rule-Based Sequential Classifiers for Web Document Prediction. *Journal of Data Mining and Knowledge Discovery.*

Zaiane, O. R., Xin, M., and Han, J. (1998). Discovering Web Access Patterns and Trends by Applying OLAP and Data Mining Technology on Web Logs. *Advances in Digital Libraries.* pp. 19 – 29.

_____. (2004). Web Server Log File Analysis – Basic. http://www.si.umich.edu/Classes/540/Readings/ServerLogFileAnalysis.htm [Date Accessed: 03-03-2004].