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

© Mohd Helmy Abd Wahab, 2004. All rights reserved
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 used 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)


iii
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

Permission to use ................................................................. i
Abstract ............................................................................... ii
Abstract (Bahasa Melayu) ...................................................... iii
Acknowledgement ................................................................. iv
Table of Contents ............................................................... v
List of Figures ....................................................................... viii
List of Tables ......................................................................... x

CHAPTER 1: INTRODUCTION

1.1 Context of the study ......................................................... 1
   1.1.1 Web Mining .......................................................... 2
   1.1.2 Association Rules ................................................... 5
   1.1.3 Web Server Log ..................................................... 5
   1.1.4 How Server Log is created ....................................... 6
1.2 Problem Statement .......................................................... 7
1.3 Objective of the Study ..................................................... 7
1.4 Scope of Project ............................................................. 7
1.5 Significance of the Study ................................................. 7
1.6 Thesis Organization ....................................................... 8

CHAPTER 2: LITERATURE REVIEW

2.1 Previous Issues and Challenges ........................................ 9
2.2 Web Mining Definition and Research Area ....................... 10
2.3 Web Mining in Education .............................................. 11
2.4 Web Usage Mining ....................................................... 11
2.5 Web Usage Mining Techniques ....................................... 12
2.6 Application Areas ....................................................... 16
CHAPTER 3: WEB SERVER LOG

3.1 Web Server Log................................................................. 19
3.2 Background........................................................................ 20
3.3 Information in Server Logs................................................ 20
  3.3.1 Access Log................................................................. 21
  3.3.2 Agent Log................................................................. 25
  3.3.3 Error Log................................................................. 26
  3.3.4 Referer Log............................................................... 27
3.4 Log File Format.................................................................. 28
  3.4.1 W3C Log File Format.................................................. 28
  3.4.2 IIS Log File Format.................................................... 29
  3.4.3 NCSA Log File Format................................................ 30

CHAPTER 4: METHODOLOGY

4.1 Introduction................................................................. 31
4.2 Raw Log File .................................................................... 32
4.3 Preprocessing................................................................... 33
4.4 Pattern Mining.............................................................. 34
4.5 Generalized Association Rules ......................................... 35
4.6 Result .............................................................................. 36

CHAPTER 5: IMPLEMENTATION

5.1 Active Server Pages ....................................................... 38
5.2 Read Server Log............................................................ 39
5.3 Transfer Server Log to Database........................................ 41
5.4 Pattern Mining Algorithm ............................................... 42
5.5 Generalized Association Rules Algorithm.......................... 47
  5.5.1 Counting Occurrence Algorithm.................................... 49
  5.5.2 Algorithm for Support.................................................. 50
CHAPTER 6: RESULT AND DISCUSSION

6.1 General Summary ......................................................... 53
6.2 Most Requested Pages ...................................................... 54
6.3 Most Downloaded Files ..................................................... 54
6.4 Most Active Countries ..................................................... 55
6.5 Most Active Search Engines .............................................. 56
6.6 Phrases ........................................................................... 56
6.7 Keyword ............................................................. 57
6.8 Support and Confidence for Structure Level 1 .......................... 58
6.9 Support and Confidence for Structure Level 2 .......................... 59
6.10 Support and Confidence for Structure Level 3 .......................... 60

CHAPTER 7: CONCLUSION AND RECOMMENDATIONS

Conclusion and Recommendation ............................................... 63

REFERENCES

References ........................................................................ 65

APPENDIXES

Appendix A: Table of Country Code Based on IP Address
Appendix B: List of Error Code
Appendix C: List of Error from Server Log Tutor.com
Appendix D: List of Popular Pages
Appendix E: List of Downloaded Files
Appendix F: List of Active Countries Access Tutor.com
Appendix G: List of Top Search Engines
Appendix H: Top Search Keyword
Appendix I: Top Search Phrases
Appendix J: Support and Confidence for Level 1, Level 2 and Level 3
# LIST OF FIGURES

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1</td>
<td>Web Mining Taxonomy</td>
<td>2</td>
</tr>
<tr>
<td>1.2</td>
<td>Web Structure Mining of Utusan Education Portal</td>
<td>4</td>
</tr>
<tr>
<td>1.3</td>
<td>Process how the Server Logs is created</td>
<td>6</td>
</tr>
<tr>
<td>3.1</td>
<td>Single Entry of Log Files from Portal Pendidikan</td>
<td>21</td>
</tr>
<tr>
<td>3.2</td>
<td>Agent Log Entry</td>
<td>25</td>
</tr>
<tr>
<td>3.3</td>
<td>Error Log File Entry</td>
<td>26</td>
</tr>
<tr>
<td>3.4</td>
<td>Referer Log Entry</td>
<td>27</td>
</tr>
<tr>
<td>3.5</td>
<td>W3C Log File Format</td>
<td>29</td>
</tr>
<tr>
<td>3.6</td>
<td>IIS Log File Format</td>
<td>30</td>
</tr>
<tr>
<td>3.7</td>
<td>NCSA Log File Format</td>
<td>30</td>
</tr>
<tr>
<td>4.1</td>
<td>Methodology of the study</td>
<td>31</td>
</tr>
<tr>
<td>4.2</td>
<td>Single Entry of Raw Log File</td>
<td>33</td>
</tr>
<tr>
<td>4.3</td>
<td>Hierarchy Structure of Portal Pendidikan Utusan</td>
<td>35</td>
</tr>
<tr>
<td>5.1</td>
<td>Method and Properties of MCSW.IISLog Class</td>
<td>39</td>
</tr>
<tr>
<td>5.2</td>
<td>Algorithm for Reading Server Logs</td>
<td>40</td>
</tr>
<tr>
<td>5.3</td>
<td>Table after data has been transferred to database</td>
<td>41</td>
</tr>
<tr>
<td>5.4</td>
<td>Algorithm transfer to database</td>
<td>41</td>
</tr>
<tr>
<td>5.5</td>
<td>Algorithm Calculating Total Hits</td>
<td>42</td>
</tr>
<tr>
<td>5.6</td>
<td>Algorithm to calculate number of page view</td>
<td>42</td>
</tr>
<tr>
<td>5.7</td>
<td>Algorithm to calculate number of download files</td>
<td>43</td>
</tr>
<tr>
<td>5.8</td>
<td>IP address conversion formula</td>
<td>44</td>
</tr>
<tr>
<td>5.9</td>
<td>IP address conversion calculation</td>
<td>44</td>
</tr>
<tr>
<td>5.10</td>
<td>META Tags</td>
<td>45</td>
</tr>
<tr>
<td>5.11</td>
<td>Referer attributes</td>
<td>45</td>
</tr>
<tr>
<td>5.12</td>
<td>Algorithm to determine the search engines, keyword and phrases</td>
<td>46</td>
</tr>
<tr>
<td>5.13</td>
<td>Algorithm to perform HTTP error analysis</td>
<td>47</td>
</tr>
<tr>
<td>5.14</td>
<td>Algorithm for data cleaning</td>
<td>48</td>
</tr>
<tr>
<td>5.15</td>
<td>Starter Prompt for Log File Simulator</td>
<td>48</td>
</tr>
<tr>
<td>5.16</td>
<td>Rule produced by the simulator</td>
<td>49</td>
</tr>
<tr>
<td>Figure 5.17</td>
<td>Counting Occurrences Algorithm</td>
<td></td>
</tr>
<tr>
<td>Figure 5.18</td>
<td>Implementation of Counting Occurrences</td>
<td></td>
</tr>
<tr>
<td>Figure 5.19</td>
<td>Support Calculation Formula</td>
<td></td>
</tr>
<tr>
<td>Figure 5.20</td>
<td>Implementation of Support Counting</td>
<td></td>
</tr>
<tr>
<td>Figure 5.21</td>
<td>Confidence Calculation Formula</td>
<td></td>
</tr>
<tr>
<td>Figure 5.22</td>
<td>Implementation of Confidence Counting</td>
<td></td>
</tr>
<tr>
<td>Figure 6.1</td>
<td>Top requested pages</td>
<td></td>
</tr>
<tr>
<td>Figure 6.2</td>
<td>Top downloaded files</td>
<td></td>
</tr>
<tr>
<td>Figure 6.3</td>
<td>Top country access to Tutor.com</td>
<td></td>
</tr>
<tr>
<td>Figure 6.4</td>
<td>Search engines used</td>
<td></td>
</tr>
<tr>
<td>Figure 6.5</td>
<td>Search phrases used in search engines</td>
<td></td>
</tr>
<tr>
<td>Figure 6.6</td>
<td>Search keyword used in search engines</td>
<td></td>
</tr>
<tr>
<td>Figure 6.7</td>
<td>Support and Confidence for level 1 for Portal Tutor.com</td>
<td></td>
</tr>
<tr>
<td>Figure 6.8</td>
<td>Support and Confidence for Question Banks</td>
<td></td>
</tr>
<tr>
<td>Figure 6.9</td>
<td>Support and Confidence for UPSR</td>
<td></td>
</tr>
<tr>
<td>Figure 6.10</td>
<td>Support and Confidence for PMR</td>
<td></td>
</tr>
<tr>
<td>Figure 6.11</td>
<td>Support and Confidence for SPM</td>
<td></td>
</tr>
<tr>
<td>Figure 6.12</td>
<td>Support and Confidence for STPM</td>
<td></td>
</tr>
<tr>
<td>Table 4.1</td>
<td>Preprocessed Log File</td>
<td>34</td>
</tr>
<tr>
<td>------------</td>
<td>---------------------------------------------</td>
<td>----</td>
</tr>
<tr>
<td>Table 6.1</td>
<td>General Summary of Server Logs</td>
<td>54</td>
</tr>
</tbody>
</table>
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


65


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


72