

ABSTRACT

The World Wide Web (WWW) is gaining popularity by providing an alternative way for individuals to get the information they require easily and quickly through the Internet. This is threatening the existence of libraries, which have been one of the main resources for information. The librarians in return are proactively planning and improving their services to meet the challenges imposed by the World Wide Web. Libraries are using the technologies of the Internet to provide new and better services to their patrons. Although the librarians are being proactive, they require information to make strategic decisions that would improve their services for their patrons. One way of getting this information is to perform data mining on the data captured in the libraries' systems. Data mining or also known as knowledge discovery from databases is a process of getting useful information or knowledge by analysing the data for patterns. The capability of data mining to solve some of the organisation's persistent and pressing issues has encouraged new information technology users to implement this system. The Extraction, Transformation and Loading (ETL) processes are required to extract, transform and load the data into a database for the data mining activities. In this dissertation, firstly, a survey was conducted on the use of data mining in the libraries of Higher Education Institutes and Universities in Malaysia. Secondly, a data extraction and data transformation application was developed for extraction, cleaning and transforming data for data mining. Finally, a prototype data mining system was implemented for the library integrating the developed data extraction and data transformation application with other suitable tools and methodology.

Acknowledgement

Firstly, I would like to thank my supervisor, Dr. Teh Ying Wah, who has given valuable direction and guidance towards completing this dissertation.

I would also like to thank the survey participants, who consist of the Higher Education Institutes' libraries and Universities' libraries in Malaysia, for their kind participation in the survey research.

I am also grateful to the Faculty of Computer Science and Information Technology staff for their support at the various level of my dissertation, especially Ms Papu and Mr. Sim.

I would also like to thank my sister, Vijaya L. Raj for her guidance and support when the going was tough.

Last but not the least; I would also like to dedicate my gratitude to my family, Mrs. R. Jayanthi, my wife, who had always been supportive, encouraging and understanding during the time I spend on this dissertation and my sons Rakesh and Shatesh for just being there.

Rabindranath Raj

TABLE OF CONTENTS

ABSTRACT	i
ACKNOWLEDGEMENTS	ii
TABLE OF CONTENTS	iii
LIST OF FIGURES	ix
LIST OF TABLES	xii
LIST OF ABBREVIATIONS	xiii
CHAPTER 1: INTRODUCTION	1
1.1 Background	1
1.2 Data Mining	4
1.3 Introduction to Extraction, Transformation and Loading (ETL) processes	6
1.4 Motivation: Why Extraction, Transformation and Loading (ETL) Processes is Required for a Data Mining System	8
1.5 Dissertation Objective and Potential Research Outcome	11
1.6 Scope of Research	12
1.7 Research Methodology	13
1.8 Content of Dissertation	15
CHAPTER 2: LITERATURE REVIEW	17
2.1 Introduction	17
2.2 Historical Perspective of Library System	19
2.2.1 Resource Information or Cataloguing	20
2.2.2 Acquisition Information	21
2.2.3 Patrons Information or Search	22

2.2.4	Circulation Information	23
2.3	Present Library Scenario	24
2.4	Data Mining and the Library	26
2.4.1	Data Mining to Improve Library Services	26
2.4.2	Data Mining as a Tool for Decision-Making by the Library Management	28
2.4.3	Data Mining for Reporting and Justification	29
2.5	Data Mining Techniques	30
2.6	Research Perspective – Data Mining for Libraries	36
2.7	ETL in Data Mining	36
2.8	The Advantages and Disadvantages of Developing an ETL Tool over Buying a Commercial ETL Tool	38
2.9	Conclusion	44
CHAPTER 3: SURVEY ON THE USE OF DATA MINING IN LIBRARIES		45
OF THE HIGHER EDUCATION INSTITUTES AND		
UNIVERSITIES IN MALAYSIA		
3.1	Introduction	45
3.2	The Survey	45
3.3	Rational of the Survey Research	47
3.4	Objectives of the Survey Research	48
3.5	Research Methodology.....	49
3.6	Population Surveyed	51
3.7	Pilot Test	52
3.8	Data Collection	53
3.9	Tools Used for Analysis	53

3.10	Survey Outcome and Analysis	53
3.10.1	High-level summary and trend analysis on whether local Higher Education Institutes or Universities libraries conduct survey on the quality of library service they provide	55
3.10.2	High-level summary and trend analysis on whether local Higher Education Institutes or Universities libraries conduct any analysis on the data captured in their database systems	56
3.10.3	High-level summary and trend analysis of the awareness of the term data mining by the librarians of the local Higher Education Institutes or Universities	57
3.10.4	High-level summary and trend analysis of the awareness of the existence of data mining tools that could be used in the library by the local Higher Education Institutes' or Universities' librarians	58
3.10.5	High-level summary and trend analysis of the usage of data mining tools by the libraries of the local Higher Education Institutes or Universities	60
3.10.6	High-level summary and trend analysis of the purpose of implementing data mining system by the libraries of the local Higher Education Institutes or Universities	61
3.10.7	High-level summary and trend analysis on whether there are any future plans of implementing a data mining system in the libraries of the local Higher Education Institutes or Universities	62
3.11	Problems Encountered	63
3.12	Discussion of the Survey	64

3.13	Conclusion	65
3.14	Limitations of the Survey	65
3.15	Appreciation	66

CHAPTER 4: DEVELOPMENT OF DATAEXTRACTION AND DATA

	TRANSFORMATION APPLICATION	67
4.1	Introduction	67
4.2	System Analysis and Requirements	69
4.3	Development Alternatives	72
4.4	System Design	73
4.5	System Implementation and Coding	77
	4.5.1 Development Environment	78
	4.5.2 Reasons for Selecting the Visual Basic 6.0 Software	79
	4.5.3 Coding	79
4.6	Testing	80
	4.6.1 Unit Testing	80
	4.6.2 System Testing	84
4.7	System Evaluation and Report	84
4.8	Problems Encountered and Solution	87
4.9	System Strength	87
4.10	System Limitation	88
4.11	Future Enhancement	88
4.12	Conclusion	89

CHAPTER 5: THE DEVELOPMENT OF A PROTOTYPE DATA MINING	
SYSTEM FOR THE LIBRARY	90
5.1 Introduction	90
5.2 Definition	91
5.3 Architecture Framework	96
5.4 Mining	100
5.4.1 Pre-Mining Activities	100
5.4.1.1 Data Preparation for Data Mining	102
5.4.1.2 Selecting and Cleaning Data	103
5.4.2 Data Mining Activities	111
5.4.3 Post Mining Activities	127
5.5 Interpretation	128
5.6 Maintenance	129
5.7 Conclusion	130
CHAPTER 6 CONCLUSION AND FUTURE DIRECTION	131
6.1 Introduction	131
6.2 Issues Encountered	132
6.3 Limitation and Weakness	134
6.4 Further Studies, Future Enhancements and Expansions	135
6.5 Significance of the Research	137
6.6 Conclusion	138
APPENDIX A: Letter to the Survey Participants	139
APPENDIX B: Survey Questionnaire	140

APPENDIX C: Project Plan	143
APPENDIX D: Project Plan's Gantt Chart	144
APPENDIX E: Source Code for Data Extraction Application	145
APPENDIX F: Source Code for Data Extraction Application	148
APPENDIX G: Application Evaluation Questionnaire	150
APPENDIX H: DAMIM Methodology	152
APPENDIX I: Review of Data Mining Methodologies	174
APPENDIX J: DAMIM Methodology Review Checklist 1	177
APPENDIX K: DAMIM Methodology Review Checklist 2	178
APPENDIX L: DAMIM Methodology Review Checklist 3	180
APPENDIX M: DAMIM Methodology Review Checklist 4	182
APPENDIX N: DAMIM Methodology Review Checklist 5	183
APPENDIX O: Change Management Form	184
APPENDIX P: Associate Rule	185
References	186

LIST OF FIGURES

Figure 3.1	Percentage of Libraries that Analysed the Data in the Database ..	56
Figure 3.2	Percentage of Awareness of Data Mining	57
Figure 3.3	Aware of Data Mining Tools for the Use of Libraries	59
Figure 3.4	Percentage of Libraries Using of Data mining Tools	60
Figure 3.5	Percentage of Libraries that Plan to Implement Data Mining Tools in the Future	62
Figure 4.1	The Linear Sequential Model	68
Figure 4.2	General System Architecture for the Data Extraction Application	74
Figure 4.3	General System Architecture for the Data Transformation Application	74
Figure 4.4	Data Extraction Application's User Interface	75
Figure 4.5	Data Transformation Application's User Interface	75
Figure 4.6	Data Extraction Application Modules	76
Figure 4.7	Data Transformation Application Modules	77
Figure 4.8	Unit Testing for Data Extraction Application	81
Figure 4.9	Unit Testing for Data Transformation Application	82
Figure 4.10	System Testing for the Data Extraction Application	83
Figure 4.11	System Testing for the Data Transformation Application	83
Figure 5.1	DAMIM Model	91
Figure 5.2:	Estimate Budget for Implementing the System	96
Figure 5.3	Architecture of the Data Mining System	97
Figure 5.4	Sample Data from the Web Log	101
Figure 5.5	Steps to Create the Database WebMining	102

Figure 5.6	Association Rule	103
Figure 5.7	Data Extraction Application	104
Figure 5.8	Extracted Data from Web Log	104
Figure 5.9	Data Transform Application	105
Figure 5.10	Transformed Data from the Extracted Data	105
Figure 5.11	Step 1 of the Data Transformation in MS SQL Database	106
Figure 5.12	Step 2 of the Data Transformation in MS SQL Database	107
Figure 5.13	Step 3, 4 & 5 of the Data Transformation in MS SQL Database ..	108
Figure 5.14	Step 6 of the Data Transformation in MS SQL Database	109
Figure 5.15	Step 7 of the Data Transformation in MS SQL Database	110
Figure 5.16	Sample of Transformed Data in the 'weblog104' Table	110
Figure 5.17	Library_fail_stream Initial Stream	111
Figure 5.18	Selected Fields for Data Mining	112
Figure 5.19	Data Mining Data	113
Figure 5.20	Library_fail_stream Model 1.....	114
Figure 5.21	Model1 Association Rule	115
Figure 5.22	Model 1 – Summary of Association Rule	116
Figure 5.23	Web node attached to the Library_fail_stream	117
Figure 5.24	Web Node Select Options	118
Figure 5.25	Web Node Display of the Patron Search Association	119
Figure 5.26	Summary of the Web Node Links Controls for the Patron Search Association	120
Figure 5.27	Web Node – Line Values above 1 Link	121
Figure 5.28	Web Node – Line Values above 2 Links	122
Figure 5.29	Web Node – Line Values Above 3	123

Figure 5.30	Web Node – Line Values Above 5	124
Figure 5.31	Web Node – Line Values Above 8	125
Figure 5.32	Web Node – Line Values Above 12	126

LIST OF TABLES

Table 3.1	Statistics of performing survey/research	55
Table 3.2	Statistics of “Percentage of Libraries that Analysed the Data in the Database”	56
Table 3.3	Statistics of “Awareness of Data Mining”	58
Table 3.4	Statistics of the Awareness of Data Mining Tools for the Use of Libraries	59
Table 3.5	Statistics of Libraries Using Data Mining Tools	60
Table 3.6	Statistics of Libraries that Plan to Implement Data Mining Tools in the Future	63
Table 4.1	Rating Scale for Questionnaire	84
Table 4.2	Questionnaire Result of the Data Extraction	85
Table 4.3	Questionnaire Result of the Data Transformation Applications ...	86

LIST OF ABBREVIATIONS

ARL	Association of Research Libraries
BI	Business Intelligence
CRISP-DM	Cross-Industry Standard Process for Data Mining
DAMIM	Data Mining Methodology
ETL	Extraction, Transformation and Loading
FTP	File Transfer Protocol
GRI	Generalised Rule Induction
ILS	Integrated Library System
KDD	Knowledge Discovery in Databases
OLAP	Online Analytical Processing
SEMMA	Sample, Explore, Modify, Model, Assess
SPSS	Statistical Package for the Social Sciences
WWW	World Wide Web