

**UNCOVERING HIDDEN INFORMATION WITHIN R&D
DEPARTMENT'S TICKET USING DATA MINING
CLUSTERING APPROACH**

AZMI BIN ABU BAKAR

UNIVERSITI UTARA MALAYSIA

2011

**UNCOVERING HIDDEN INFORMATION WITHIN R&D
DEPARTMENT'S TICKET USING DATA MINING
CLUSTERING APPROACH**

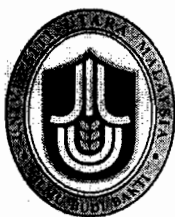
**A thesis submitted to the Faculty of Information Technology
in partial fulfillment of the requirement for the
degree Master of Science (Information Technology)**

Universiti Utara Malaysia

By

AZMI BIN ABU BAKAR

@ Azmi Bin Abu Bakar, 2011. All rights reserved.



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, certifies that)

AZMI BIN ABU BAKAR
(802332)

calon untuk Ijazah
(candidate for the degree of) **MSc. (Information Technology)**

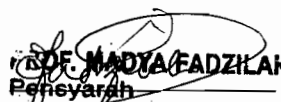
telah mengemukakan kertas projek yang bertajuk
(has presented his/ her project of the following title)

UNCOVERING HIDDEN INFORMATION WITHIN R&D DEPARTMENT
TICKET USING DATA MINING CLUSTERING APPROACH

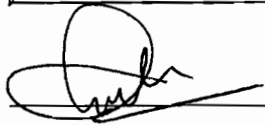
seperti yang tercatat di muka surat tajuk dan kulit kertas projek
(as it appears on the title page and front cover of project)

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

Nama Penyelia
(Name of Supervisor) : **ASSOC. PROF. FADZILAH SIRAJ**

Tandatangan
(Signature) :  Tarikh (Date) : 27/2/2011

Nama Penilai
(Name of Evaluator) : **MR. TUAN ZALIZAM TUAN MUDA**

Tandatangan
(Signature) :  Tarikh (Date) : 27/2/2011

TUAN ZALIZAM BIN TUAN MUDA
Lecturer
Information Technology Building
College Of Arts & Sciences
Universiti Utara Malaysia

PERMISSION TO USE

In presenting this project in partial fulfillment 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, in her absence, by the Dean of the Faculty of Information Technology. It is understood that any copying or publication or use of this project or parts thereof for financial gain should 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 use of material in this project, in whole or in part should be addressed to:

Dean of the Faculty of Information Technology

Universiti Utara Malaysia

06010 UUM Sintok

Kedah Darul Aman

ABSTRAK

Syarikat MP bahagian R&D Pulau Pinang menerima tiket setiap hari (setiap tiket yang dikemukakan mewakili satu isu yang dihadapi oleh pekerjaanya) dari pekerjaanya yang memohon bantuan sokongan. Jumlah tahunan tiket yang dikemukakan bertambah dari tahun 2007 sehingga 2009. Bertambahnya isu bermakna bertambahnya aktiviti sokongan dibuat bagi menyelesaikan isu-isu tambahan yang dihadapi. Secara langsung ini bermakna bertambahnya kos operasi. Projek ini adalah bertujuan untuk membuat analisa dengan menggunakan cara yang di deskripsi oleh teknik *Data Mining*, iaitu Clustering analysis. Maklumat tersembunyi dan sebab utama berlakunya pertambahan isu-isu tahunan dapat di bentangkan. Hasil keputusan dari analisa ini akan digunakan bagi mengolah satu kerangka ataupun penyelesaian bagi memperbaiki keadaan tersebut dan menstabilkan jumlah tiket-tiket yang dikemukakan.

Dalam kajian ini data yang diperolehi di clusterkan dengan menggunakan dua teknik data mining yang berbeza iaitu K-Mean dan Kohonen Network. Kemudian perbandingan keatas cluster-cluster yang terhasil dibuat dan dinilai dengan menggunakan Multinomial Logistic Regression dan Neural Network:MLP. Hasil keputusan menyerlahkan sebab utama (root cause) yang menyebabkan tiket-tiket dikemukakan. Maklumat ini akan digunakan oleh MP Company untuk menghasilkan kerangka ataupun kaedah penyelesaian bagi aplikasi.

ABSTRACT

MP Company's R&D department in Penang received daily submission tickets (this represent issues raised by its staffs) from it staff requesting for support. Number of annual tickets submission increases from year 2007 until 2009. Increase of issues means that increase of support activities in order to resolve these extra issues. Directly this will increase the cost of operation. This project will undergo analysis which prescribes in one of data mining technique called Clustering analysis. Hidden information and major root cause of the increase issues is expected to be unveiled. Result of this analysis can be used to generate framework or solution to improve the situation and stabilized the number of tickets submission

In this study the data extracted is clustered using two different types of data mining techniques i.e. K-Means and Kohonen Network. Later the clustered produced is compared and evaluated using Multinomial Logistic Regression and Neural Network: MLP. The result produced then reveals the biggest root caused of issue or problems that eventually triggered the ticket being submitted. This knowledge will be used by MP Company to further produce the framework or solution model for implementation.

ACKNOWLEDGEMENT

I would like to express my utmost gratitude to my supervisor Associate Professor Fadzilah Binti Siraj, whom with abundance patients, has taught and navigate me in undergoing this whole thesis.

I would also like to thanks my family who inspired, encouraged and fully supported me morally and spiritually along the way.

Last but not least, to my course mates and my friends who have helped me directly or indirectly in ensuring the completion of this research paper and tasks.

TABLE OF CONTENT

PERMISSION TO USE	i
ABSTRAK	ii
ABSTRACT	iii
ACKNOWLEDGEMENT	iv
TABLE OF CONTENTS	v
LIST OF TABLES	viii
LIST OF FIGURES	ix
LIST OF ABBREVIATIONS	xi
CHAPTER 1: INTRODUCTION 1	1
1.1 Problem Statement	2
1.2 Research Objective	3
1.3 Research Scope	4
1.4 Significance of Study	4
1.5 Summary	4
CHAPTER 2: LITERATURE REVIEW	5
2.1 Data Mining	5
2.2 Knowledge Discovery in Data Mining	6
2.3 CRISP-DM	7
2.3.1 Business Understanding	9
2.3.2 Data Understanding	9
2.3.3 Data preparation	10
2.3.4 Modeling	10

a)	Cluster analysis	11
b)	Two-step	11
c)	K-Means clustering	11
d)	Kohonen Neural Network	12
2.3.5	Evaluation	13
a)	Multinomial Logistic regression (MLR)	13
b)	Neural Network	13
2.3.6	Deployment	14
2.4	Data Mining Hybrid Model	15

CHAPTER 3: METHODOLOGY **17**

3.1.	CRISP-DM	17
3.1.1	Business understanding	18
3.1.2	Data understanding	19
3.1.3	Data preparation	21
3.1.4	Modeling	23
3.1.5	Evaluation	30
3.1.6	Deployment	31

CHAPTER 4: RESULT **32**

4.1	K-Means Clustering	32
4.2	Neural Network: Kohonen Network	35
4.3	Multinomial Logistic Regression	38
4.3.1	Using all variable as covariate	40
4.3.2	Using OPERATION2+PRODUCT3 as covariate	43

a)	For Cluster 2	43
b)	For Cluster 3	43
c)	For Cluster 4	43
d)	Model generated	44
4.4	Neural Network: MLP	45
4.4.1	Parameters set 1 (This is SPSS default setting)	45
4.4.2	Parameters set 2	47

CHAPTER 5: CONCLUSION 49

5.1	Conclusion	49
-----	------------	----

5.2	Recommendation	50
-----	----------------	----

REFERENCE

APPENDICES

Appendix A: DATA MAPPING ON VARIABLES VALUE SUBSTITUTION

Appendix B: SCREENSHOT OF THE RAW DATA

LIST OF TABLES

Tables	Title	Page
Table 3.1	Frequency of ticket by location of submitters	21
Table 3.2	Result of auto-clustering via Two-Step	26
Table 3.3	Show the assignment of cluster to data using K-Means clustering (cut-off-at center)	29
Table 4.1	Criteria of cluster (K-Means)	33
Table 4.2	Spearman Correlation (K-Means)	34
Table 4.3	Criteria of cluster (Kohonen)	36
Table 4.4	Spearman Correlation (Kohonen)	37
Table 4.5	Result from frequency test.	38
Table 4.6	Result of beta value from MLR	40
Table 4.7	Classification table and prediction	42
Table 4.8	Result of beta value from MLR	44
Table 4.9	Classification table and prediction	45
Table 4.10	Classification table and prediction (Parameters set 1)	46
Table 4.11	Classification table and prediction (Parameters set 2)	48

LIST OF FIGURES

Figures	Title	Page
Figure 1.1	MP Company ticketing process flow	3
Figure 2.1	Process of Knowledge Discovery from Data Mining. (Han & Kamber, 2006)	7
Figure 2.2	Generic tasks (bold) and outputs (italic) of the CRISP-DM reference model	9
Figure 2.3	Common model of Neural Network (source NeuroDimension 2010)	14
Figure 2.4	Hybrid Model source Cios et. al, 2007	16
Figure 3.1	Phases of CRISP-DM Reference Model	18
Figure 3.2	Adapted from Porter (1979)	19
Figure 3.3	Graph of percentage of ticket submission by location	22
Figure 3.4	Data set loaded into SPSS application	24
Figure 3.5	Menu option chosen for two-step	24
Figure 3.6	Variable selected for Two-Step processing	25
Figure 3.7	The option selected in the <i>Plot</i> option	25
Figure 3.8	Menu selections on K-Means clustering	27
Figure 3.9	Increase the number of iteration	28
Figure 3.10	The option selected K-means clustering	28
Figure 4.1	Cluster Frequency percentage in graphical mode	34
Figure 4.2	Percentage of Cluster members via Kohonen Network	37
Figure 4.3	NN network generated with 3 neuron of hidden layer and 2	47

bias neurons

Figure 4.4 NN network generated with 4 neuron of hidden layer and 2

bias neurons

48

LIST OF ABBREVIATIONS

CRISP-DM	Cross-Industry Standard Process for Data Mining
DM	Data Mining
KDDM	Knowledge Discovery in Data Mining
KPI	Key Performance Index
LR	Logistic Regression
MLP	Multi-Layer Perceptron
MLR	Multinomial Logistic Regression
MP	The company whose data being analyzed
NN	Neural Network
R&D	Research and Development
SME	Small Medium Enterprise

CHAPTER 1

INTRODUCTION

In MP Company, for the Research and Development (R&D) division, a ticketing system is implemented. It is a system for user to submit a ticket from R&D staffs that use standard MP tools and applications, from hereon it referred to as users, who have request or issues their needs to be resolved internally. The issues can be ranging from *forgetting the password* up to *system development failure*.

Generically, a ticket can be viewed as a single complaint or a single request from a user to the supporting team, requesting work to be done to resolve the request or issue. Once the request is fulfilled or the issue is resolved, the responsible team will input feedback in the ticket and close the ticket. The feedback may contain work progress, details of works and status of ticket; either in progress, on hold, closed and etc.

The ticket details are stored on a centralized location in Oracle database. In this analysis the ticket extracted is filtered on tickets that have been resolved by the IT team for MP R&D in Penang. The ticket is also being taken from time to time as key

The contents of
the thesis is for
internal user
only

REFERENCE

- BMC Software (2011). Retrieved 21 Feb 2011 from <http://www.bmc.com/products>
- Bošnjak, Z., Grljević, O, & Bošnjak, S. (2009). *CRISP-DM as a Framework for Discovering Knowledge in Small and Medium Sized Enterprises' Data. 5th International Symposium on Applied Computational Intelligence and Informatics held on 28 May–29 May 2009 at Timișoara, Romania*
- Catley, C., Smith, K., McGregor, C., & Tracy, M. (2009). *Extending CRISP-DM to Incorporate Temporal Data Mining of Multidimensional Medical Data Streams: A Neonatal Intensive Care Unit Case Study*. University of Ontario Institute of Technology, Oshawa, Canada
- Chan, Y.H. (2005). *Basic Statistics For Doctors: Biostatistics 305. Multinomial logistic regression*.(CME Article). Singapore Med J.
- Choinski, M., & Chudziak, J. A. (2009) Proceedings of the International Multiconference on Computer Science and Information Technology: *Ontological Learning Assistant for Knowledge Discovery and Data Mining*. pp. 147–155
- Cios, K. J., Pedrycz, W., & Swiniarski, W. (2007). *Data Mining: A Knowledge Discovery Approach*, New York: Springer Science + Business Media, LLC
- Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Thomas, R., Shearer, C., & Wirth, R., (2000). *CRISP-DM 1.0 Step-by-step data mining guide*.
- Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Thomas, R., Shearer, C., & Wirth, R., (1999). *The CRISP-DM Process Model*. CRISP-DM Discussion Paper.
- Chen, W.B., and Zhang C., (2006). *A Robust Method for Biological Sequence Clustering*. Alabama: University of Alabama
- Crabtree, D., Andreae, P. & Gao, X., (2007). *Understanding Query Aspects with applications to Interactive Query Expansion*, New Zealand: Victoria University of Wellington
- Gashi, I., Stankovic V., Leita, C. and Thonnard, O., (2009). *An Experimental Study of Diversity with Off-The-Shelf AntiVirus Engines*, London: City University London
- Ghaderi, S.F., Azadeh, A., & Keyno, H. S. (2007). *Forecasting electricity consumption by separating the periodic variable and decompositions the pattern*, Iran: University of Tehran

- Han, J., & Kamber, M. (2006). *Data Mining: Concepts and Technique* (2nd ed.). San Francisco: Morgan Kaufmann Publisher
- Harun, N., Dlay, S.S., & Woo, W.L. (2010). *Performance of Keystroke Biometrics Authentication System Using Multilayer Perceptron Neural Network (MLP NN)*. United Kingdom: Newcastle University.
- Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference and Prediction*. New York: Springer Science + Business Media, LLC
- Heaton, J. (2005). *Introduction to Neural Network with Java*. Chesterfield: Heaton Research Inc.
- Hu, J., Xiong, C., Shu, J., Zhou, X. & Zhu, J. (2009). *A Novel Text Clustering Method Based on TGSOM and Fuzzy K-Means*, First International Workshop on Education Technology and Computer Science
- IBM Corp.(2010). IBM® Rational® ClearCase® offers complete software configuration management. In *Rational ClearCase*. Retrieve Feb 26, 2011 from <http://www-01.ibm.com/software/awdtools/clearcase/>
- Matteucci, M. (n.d.). K-Means Clustering. In *A Tutorial on Clustering Algorithm*. Retrieved Feb 25, 2011 from http://home.dei.polimi.it/matteucc/Clustering/tutorial_html/kmeans.html
- Norsaadah Zakaria, Jamil Salleh, Mohd Nasir Taib, Tan, Y. Y. & Wah, Y.B., (2008). *Using Data Mining Technique to Explore Anthropometric Data towards the development of Sizing System*. Malaysia: Universiti Teknologi Mara
- NeuroDimension Inc (2010). What is Neural Network?. In NeuroSolution. Retrieved Feb 25, 2011, from <http://www.nd.com/neurosolutions/products/ns/whatisNN.html>
- Pan, D. (2009). *A Formal Framework for Data Mining Process Model*, Second Asia-Pacific Conference on Computational Intelligence and Industrial Applications
- Ping, O. A. (2009). *Analysis of Bankruptcy using data mining approach*. Unpublished master's thesis, Universiti Utara Malaysia, Kedah.
- Fadzilah Siraj & Abdoulha, M.A.(2010). Mining Enrolment Data Using Predictive and Descriptive Approaches. In K. Funatsu & K. Hasegawa (Eds.), *Knowledge-oriented application in Data mining*.(pp. 53-72). Croatia: InTech
- Flom, P. (n,d). Multinomial Logistic Regression. In Statistical Analysis Consulting. Retrieved Feb 25, 2011, from

<http://www.statisticalanalysisconsulting.com/statistical-analysis-for-dissertations/research-methods/multinomial-logistic-regression/>

- Porter, Michael E. (1979). *How competitive forces shape strategy*. Harvard business review, 57(2): 137-145.
- Rahm, E & Do, H. H. (Dec 2000). *Data Cleaning: Problems and Current Approaches*, Germany: University of Leipzig
- Selbaş, R., Şencan, A. & Küçüksille, E.U. (2010). Data Mining Methods for Energy System Applications. In K. Funatsu & K. Hasegawa (Eds.), *Knowledge-oriented application in Data mining*. (pp. 339-352). Croatia: InTech
- Shearer, C. (2000). The CRISP-DM model: the new blueprint for data mining, *Journal of Data Warehousing*, vol. 5, pp. 13-19
- SPSS Inc. (2003). *SPSS Training: Introduction to Clementine*. Retrieved 29 Jan 2011 from <http://homepage.univie.ac.at/marcus.hudec/Lehre/WS%202006/Methoden%20DA/IntroClem.pdf>
- SPSS Inc. (2007). *SPSS Neural Network 16.0*. Retrieved 20 Feb 2011 from http://www.uni-muenster.de/imperia/md/content/ziv/service/software/spss/handbuecher/englisch/spss_neural_network_16.0.pdf
- Sumathi, S., & Sivanandam, S. N. (2006). *Introduction to Data Mining and its Applications*, New York: Springer-Verlog Berlin Heidelberg
- Tan, P.N., Steinbach, M. & Kumar, V., (2006). *Introduction to Data Mining*, Addison-Wesley
- UCLA: Academic Technology Services, Statistical Consulting Group (n.d.). *Annotated SPSS Output Multinomial Logistic Regression*, Retrieved 20 Feb 2011 from <http://www.ats.ucla.edu/stat/spss/output/mlogit.htm>
- Webb, K. (2010). *Cluster Analysis*. Retrieved 23 Feb 2011, from http://www.cob.sjsu.edu/webb_k/B231A/ClusterR.doc
- Wikipedia (2011). Self- Organizing map. Retrieved 25 Feb 2011, from http://en.wikipedia.org/wiki/Self-organizing_map
- Yuan, W., & Yihua, Z., (2009). *Research on Classification and Subdivision Model of Telecom Rural Channel Based on Clustering Analysis*. International Conference on Information Management, Innovation Management and Industrial Engineering
- Yun, C.H., Chuang, K.T., & Chen, M.S. (n.d). *An Efficient Clustering Algorithm for Market Basket Data Based on Small Large Ratios*, Taiwan : National Taiwan University

- Zeng, H., & Pan, D. (2010). *A Knowledge Discovery and Data Mining Process Model in E-Marketing*. 8th World Congress on Intelligent Control and Automation held on July 6-9 2010 at Jinan, China
- Zhou, J. (2004). *Comparing Regularized B-spline Neural Network, Multilayer Perceptron and Boosted-CART on Two Problems of Heart Arrhythmia Diagnosis*. Proceedings of the 26th Annual International Conference of the IEEE EMBS, San Francisco