AN INFORMATION RETRIEVAL ALGORITHM TO EXTRACT INFLUENTIAL FACTORS

A project submitted to Dean of Research and Postgraduate Studies Office in partial Fulfillment of the requirement for the degree Master of Science (Information Technology) Universiti Utara Malaysia

By
NABILAH FILZAH BINTI MOHD RADZUAN
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 purpose may be granted by my supervisor(s) or, in their absence by the Dean of Postgraduate and Research. It is understood that any copying or publication or use of this project 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.

Requests 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 Research and Postgraduate Studies
College of Arts and Sciences
Universiti Utara Malaysia
06010 UUM Sintok
Kedah Darul Aman
Malaysia
ABSTRAK (BAHASA MALAYSIA)


III
ABSTRACT (ENGLISH)

Past literatures showed that there are many factors that can be used to assess company’s performance but only a limited number of factors are needed to efficiently assess its performance. The aim of the study is to develop an algorithm that can extract a minimum set of factors that can be used to assess companies’ performances. Stock price was used as the dependent factor. The factors extracted are known as influential factors because these factors were found to have strong influence on the stock price. The objectives of the study were to obtain a comprehensive influential factors from past literatures, develop an extraction algorithm that can identify influential factors, and present factors that influenced companies’ stock prices. Data consisted of financial factors that were obtained from financial documents of distressed companies and non-distressed companies listed on a stock exchange. The extraction algorithm was developed and implemented using Matlab programming language. Results showed that out of 33 factors, 5 factors were found to be the minimum set needed to assess the companies’ performances. These were debt, investment, total asset, asset turnover, and working capital. The algorithm were tested on other dataset and results produced more than 70 percent of positive feedback. This indicates that the algorithm was able to produce a good model. The extraction algorithm developed showed that influential factors produced could be used as guideline for companies to monitor and strategize ways for business improvement.
ACKNOWLEDGEMENT

Firstly, praise to Allah S.W.T. for guiding and blessing with perseverance and strength to complete the project. Apart from the efforts of me, the success of the project depends largely on the encouragement and guidelines of many others. I take this opportunity to express my gratitude to the people who have been instrumental in the successful completion of this project. The special thank goes to my helpful supervisor Dr. Faudziah Ahmad and Prof. Ku Ruhana Ku Mahamud. I can’t say thank you enough for his tremendous support and help. Without her encouragement and guidance this project would not have materialized. My grateful thanks also go to my parents and siblings, who held faith in me and pushed me to succeed. A big contribution and supported from you is very great indeed. Special thanks also go to my friends those who supported and motivated me during the project completion was vital for the success of the project. Last but not least, I would like to thank to all University Utara Malaysia management especially College of Arts and Sciences staff and those who involved directly or indirectly in the project. May Allah bless all of you.
TABLE OF CONTENTS

PERMISSION TO USE II
ABSTRACT (BAHASA MALAYSIA) III
ABSTRACT (ENGLISH) IV
ACKNOWLEDGMENTS V
LIST OF TABLE IX
LIST OF FIGURES X
LIST OF ABBREVIATION XI
CHAPTER ONE: INTRODUCTION
1.1 Background 1
1.2 Problem statement 4
1.3 Project’s objective 5
1.4 Scope of research 5
1.5 Contribution of research 5
1.6 Summary 5

CHAPTER TWO: LITERATURE REVIEW
2.1 Introduction 6
2.2 Influential factors 6
2.2.1 Definition of influential factors 6
2.2.2 Influential factors from previous research 7
2.2.3 Stock price 12
2.3 Data mining and KDD approach 13
2.4 Algorithms for extraction 15
2.5 Validation of Algorithms 17
2.6 Knowledge theory in business 17
2.7 Summary 18
CHAPTER THREE: RESEARCH METHODOLOGY
3.1 Introduction 19
3.2 Research approach 19
   3.2.1 Phase I - Data Selection and Preprocessing 21
   3.2.2 Phase II - Data Discretization 22
   3.2.3 Phase III - Algorithm Construction 23
   3.2.4 Phase IV - Knowledge Extraction 23
3.3 Summary 24

CHAPTER FOUR: RESULT AND DISCUSSION
4.1 Introduction 25
4.2 Findings and Results 25
   4.2.1 Phase I - Data Selection and Preprocessing 25
   4.2.2 Phase II - Data Discretization 31
   4.2.3 Phase III - Algorithm Construction 34
   4.2.4 Phase IV - Knowledge Extraction 34
4.3 Summary 34

CHAPTER FIVE: EXTRACTION ALGORITHMS
5.1 Introduction 35
5.2 Extraction Algorithm 35
5.3 Step for the whole extraction algorithm 39
5.4 Implementation of the extraction algorithm 41
5.5 Knowledge Extraction 61
5.6 Summary 61

CHAPTER SIX: CONCLUSION AND RECOMMENDATION FOR FUTURE STUDY
6.1 Introduction 62
6.2  Discussion of findings  62

6.3  Limitations  64

6.4  Contributions  64

6.5  Future works  64

REFERENCES  65
APPENDIC A  69
APPENDIC B  73
APPENDIC C  74
APPENDIC D  75
LIST OF TABLES

Table 2.1: List of factors ........................................... 7
Table 2.2: List of dependent factors ......................... 9
Table 2.3: List of independent factors .................... 10
Table 2.4: List of control factors .......................... 10
Table 2.5: List of exogenous factors ....................... 11
Table 2.6: List of endogenous factors .................... 11
Table 4.1: The list of distressed companies ..................... 26
Table 4.2: The list of non-distressed companies ............. 27
Table 4.3: List of 18 factors ..................................... 28
Table 4.4: Code of distribution fitting ....................... 36
LIST OF FIGURES

Figure 1.1: The relationship between specific algorithms and business analytical problem areas. (Adapted from Nisbet et al.(2009)) 3

Figure 2.1: The phase of CRISP-DM (Adapted from Chapman et al., 2000) 13

Figure 2.2: The overall KDD process (Adapted from Fayyad et al. (1996)) 14

Figure 3.1: Research approach 19

Figure 3.2: Activites involve in each phase 20

Figure 3.3: The flow of phase I 21

Figure 3.4: The flow of phase II 22

Figure 3.5: The flow of phase III 23

Figure 3.6: The flow of phase IV 23

Figure 4.1: Binarization process 30

Figure 4.2: Obstime (original) data and fit 1 (discretized) data 31

Figure 4.3: Value of x and the discretized value (fit 1) 32

Figure 4.4: Mean and variance of dataset 32

Figure 4.5: Code of distribution fitting (The MathWorks,Inc(2009)) 33

Figure 4.6: Partitioning of dataset 35

Figure 5.1: The main call code for algorithms 41

Figure 5.2: The main code for SBS model 42

Figure 5.3: The main code for SFS model 50

Figure 5.4: The main code for reliefF model 56

Figure 5.5: Main interface 60

Figure 5.6: Result from using SFS factors selection method 61
<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>KDD</td>
<td>KNOWLEDGE DISCOVERING IN DATABASE</td>
</tr>
<tr>
<td>EM</td>
<td>EXPECTATION MAXIMIZATION</td>
</tr>
<tr>
<td>CCR CURVE</td>
<td>CORRELATED COMPONENT REGRESSION CURVE</td>
</tr>
</tbody>
</table>
CHAPTER ONE

INTRODUCTION

1.1 Background

Influential factors are significant factors which can bring impact to the environment of studies. According to Ouwen et al. (2001), influential is when a data structure has a large impact on the estimated model parameters and their characteristics are investigated to know how much they influenced the environment. The influential factors can be captured from large or small datasets or databases using some methods and represent useful information to a company. Gray (1989) stated that the influential data present important clues about the model and process under study. One approach that can detect influential data structures is to compare the estimates of the model parameters based on the sample with and without the data structures of interest using an algorithm (Ouwens et al., 2001). According to Chatterjee and Hadi (1986), influential factors can be classified into five groups which include measures based on residuals, prediction matrix, volume of confidence ellipsoids, influence functions, and partial influence. In this research, an algorithm has been developed to identify influential factors that are based on influence functions. The identified influential factors can be used as guidelines or a basis for companies to strategize their plans for improving business performances.

In accessing a company’s performance, many factors are considered. These include independent factors, dependent factors, and control factors (Maiga and Jacobs, 2003).
The contents of the thesis is for internal user only
References


Oracle. (2005). *Oracle @ Data Mining Concepts. 11g Release 1(11.1)*. Part number B28129-04

Osuna, R. (2002). Pattern Analysis. *CSCE 666, CSE@TAMU*.


