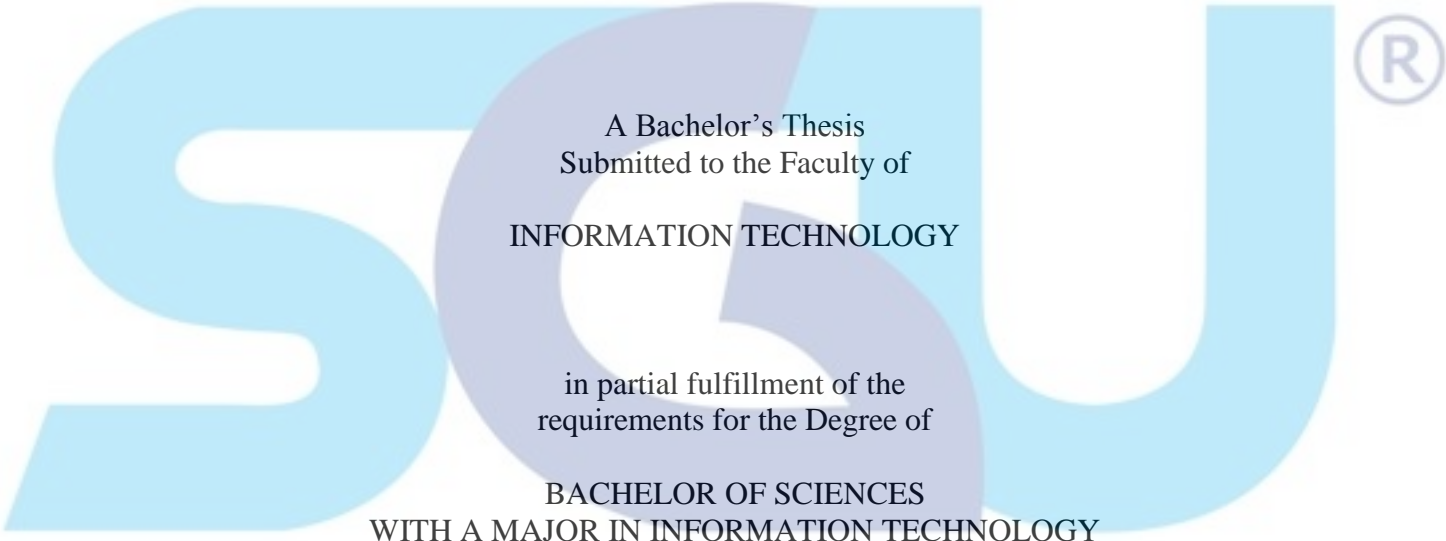


**INVESTIGATING ON HOW BETTER DATA QUALITY
AFFECTING THE HIERARCHICAL CLUSTERING PROCESS
TIME**

By

Michael



A Bachelor's Thesis
Submitted to the Faculty of
INFORMATION TECHNOLOGY

in partial fulfillment of the
requirements for the Degree of
BACHELOR OF SCIENCES
WITH A MAJOR IN INFORMATION TECHNOLOGY

SWISS GERMAN UNIVERSITY

SWISS GERMAN UNIVERSITY
Campus German Centre
Bumi Serpong Damai – 15321
Island of Java, Indonesia
www.sgu.ac.id

July 2010

Revision after the Thesis Defense on 29 July 2010

STATEMENT BY THE AUTHOR

I hereby declare that this submission is my own work and to the best of my knowledge, it contains no material previously published or written by another person, not material which to a substantial extent has been accepted for the award of many other degree or diploma at any educational institution, except where due acknowledgement is made in the thesis.

Michael

Date

Approved by:

Ir. Basuki Setio Msc

Date

Chairman of the Examination Steering Committee

Date

Michael

ABSTRACT

INVESTIGATING ON HOW BETTER DATA QUALITY AFFECTING THE HIERARCHICAL CLUSTERING PROCESS TIME

By

Michael

SWISS GERMAN UNIVERSITY

Bumi Serpong Damai

Ir. Basuki Setio Msc

Hierarchical Clustering is one of many clustering methods used to exploring the relationship in statistical data. It cluster data based on distance, similarities and correlation. The result will be shown as Dendogram.

When developing a hierarchical clustering application, most of the problems occur on how you manage the data and how you process it so it will not use too many resources in your computer in order to reduce the running time.

The application will be developed using C# with .Net 3.5 frameworks in Visual Studio 2008 as our IDE (Integrated Development Environment) and Windows Vista as the operating system.

Keywords – Hierarchical Clustering, Data, Dendogram, Running time

DEDICATION

I dedicate this thesis to both of my parent, Hendra Satyo and Dian Budiman.



ACKNOWLEDGMENTS

First of all, I want to give the utmost thanks to God for His help and blessing from the start until I finish writing this thesis.

I would also like to give special thanks to Mr. Basuki Setio for he has helped and support me in developed this application from a scratch with his advice and guidance to make this application run well.

In addition, I also give thanks to Mr. Anto Satriyo Nugroho for helping me creating clear understanding about how hierarchical clustering works and how we implemented it.

Finally, I would like to thank my friends who always support me and helping me during my Thesis work.



SWISS GERMAN UNIVERSITY

TABLE OF CONTENTS

STATEMENT BY THE AUTHOR.....	2
ABSTRACT.....	3
DEDICATION.....	4
ACKNOWLEDGMENTS	5
CHAPTER 1 – INTRODUCTION	12
1.1 Background.....	12
1.2 Objective.....	13
1.3 Research problems.....	13
1.4 Significance of study.....	13
1.5 Scope of analysis.....	13
1.6 Proposed work	14
1.7 Result	14
1.8 Thesis organization	14
CHAPTER 2 – LITERATURE REVIEW	15
2.1 Data mining.....	15
2.1.1 What is data mining.....	15
2.1.2 Data mining evolution.....	16
2.1.3 Motivating Challenges in Data Mining	17
2.1.3.1 Scalability	17
2.1.3.2 High dimensionality	17
2.1.3.3 Heterogeneous and Complex data	17
2.1.3.4 Data ownership and distribution.....	18
2.1.3.5 Non-Traditional analysis.....	18
2.1.4 Data Mining Tasks	18
2.1.4.1 Predictive modeling	19
2.1.4.2 Association analysis.....	19
2.1.4.3 Cluster analysis	19
2.1.4.4 Anomaly detection.....	19

2.2	Cluster analysis	19
2.2.1	Purpose	20
2.2.1.1	Clustering for understanding.....	20
2.2.1.2	Clustering for utility.....	20
2.2.2	Different types of clustering.....	21
2.2.2.1	Hierarchical versus partition.....	21
2.2.2.2	Exclusive versus overlapping versus fuzzy	21
2.2.2.3	Complete versus partial.....	21
2.2.3	Different types of cluster.....	21
2.2.3.1	Well-separated	22
2.2.3.2	Prototype-based.....	22
2.2.3.3	Graph-based	22
2.2.3.4	Density-based.....	22
2.2.3.5	Shared property	23
2.3	Hierarchical Clustering	23
2.3.1	Basic agglomerative hierarchical clustering algorithm	23
2.3.2	Specific techniques.....	26
2.3.2.1	Single link or MIN	27
2.3.2.2	Complete link or MAX.....	30
2.3.2.3	Group average.....	33
2.4	Weka	35
2.5	C# with .NET framework 3.5	37
CHAPTER 3 – METHODOLOGY		38
3.1	Dataset.....	38
3.2	Preprocessing	38
3.3	Proximity matrix	41
3.4	Clustering process.....	42
3.4.1	Single link.....	43
CHAPTER 4 – RESULT AND DISCUSSION		49
4.1	Preprocessing algorithm.....	49
4.2	Proximity matrix algorithm.....	63
4.3	Hierarchical Clustering algorithm (Single link)	69

CHAPTER 5 – CONCLUSION AND RECOMMENDATION	78
5.1 Conclusion	78
5.2 Recommendation	81
REFERENCES	83
APPENDICES	84
CURRICULUM VITAE.....	126

