



TABLE OF CONTENTS

APPROVAL PAGE	iii
STATEMENT.....	v
FOREWORD.....	vi
LIST OF FIGURES	x
LIST OF TABLES	xii
ABSTRACT	xiii
CHAPTER I INTRODUCTION.....	1
1.1 Research Background.....	1
1.2 Research Problem	3
1.3 Research Scope.....	3
1.4 Research Objective	3
1.5 Research Advantages.....	4
1.6 Research Methodology	4
CHAPTER II LITERATURE REVIEW	6
CHAPTER III BASIC THEORY	9
3.1 IT Audit	9
3.1.1 Audit Evidence	9
3.2 Data Mining.....	9
3.2.1 Text Mining.....	10
3.3 Document Collection.....	11
3.3.1 Features of Document.....	11
3.4 Information Retrieval.....	12
3.2.1 Searching by Query	13
3.5 Search Engine	13
3.6 Preprocessing Document.....	14
3.6.1 Tokenization.....	14
3.6.2 Stopword Removal	14
3.6.3 Stemming	14
3.7 Document Clustering.....	15
3.8 LINGO Algorithm	16
3.8.1 Preprocessing Document in LINGO	17
3.8.2 Feature Extraction	18
3.8.3 Clustering Label Induction	19
3.8.4 Cluster Content Discovery.....	19
3.8.5 Final Cluster Formation.....	20
3.9 Comparison Algorithm.....	20
3.9.1 Suffix Tree Algorithm	20
3.9.2 Bisecting K-Means Algorithm.....	22
3.10 Carrot2	24
3.11 Apache Solr	25
3.12 Carrot2 Visualization	26
3.12.1 Folder List.....	27
3.12.2 Foamtrees.....	28
3.12.3 Circles	29
3.13 Evaluation Performance.....	29



CHAPTER IV ANALYSIS AND DESIGN	31
4.1 System Analysis.....	31
4.2 Description of System	31
4.2.1 System Requirement Analysis	33
4.3 Data Preparation.....	33
4.4 Data Preprocessing	36
4.4.1 Case Folding.....	38
4.4.2 Non-alphanumeric Character Removal.....	39
4.4.3 Tokenization	40
4.4.4 Stopword Removal	41
4.4.5 Stemming	43
4.5 Indexing Process	46
4.6 Clustering Algorithm Design	49
4.6.1 LINGO Algorithm	52
4.6.2 Suffix Tree Clustering Algorithm	56
4.6.3 Bisecting K-Means	58
4.7 Carrot2	60
4.8 Visualization Schema.....	60
4.9 Evaluation Performance.....	61
CHAPTER V IMPLEMENTATION	62
5.1 Implementation Tools.....	62
5.2 System Implementation	62
5.2.1 Data Preparation	62
5.2.2 Data Preprocessing	64
5.3 Indexing Data.....	66
5.3.1 Schema Design	69
5.4 Clustering Implementation	71
5.5 Carrot2 Visualization	73
5.5.1 Web Application.....	74
5.5.2 Workbench Application.....	76
CHAPTER VI RESULT AND DISCUSSION.....	78
6.1 Data Preprocessing Result	78
6.2 Indexing Experiment	79
6.3 Clustering Experiment	82
6.4 Visualization Result	83
6.4.1 Visualization Result by using LINGO	84
6.5 Evaluation Performance.....	90
6.5.1 Search Result Evaluation	90
6.5.2 Clustering Result Evaluation	93
CHAPTER VII CONCLUSION AND FUTURE WORKS.....	99
7.1 Conclusion	99
7.2 Future Works.....	99
REFERENCES	101
ATTACHMENT.....	104