

# **Investigating the Applicability of Partition–Based Clustering for Sinhala Documents**

D.A.Meedeniya



University of Moratuwa, Sri Lanka.  
Electronic Theses & Dissertations  
[www.lib.mrt.ac.lk](http://www.lib.mrt.ac.lk)

This dissertation was submitted to the  
Department of Computer Science and Engineering  
of the University of Moratuwa Sri Lanka  
in partial fulfilment of the requirement for the  
Degree of MSc in Computer Science specializing in Software Architecture

Department of Computer Science and Engineering  
University of Moratuwa.  
Sri Lanka  
December – 2008

## Declaration

“I certify that this dissertation does not incorporate without acknowledgement any material previously submitted for a degree or diploma in any university; and to the best of my knowledge and belief it does not contain any material previously published or written by another person except where due reference is made in the text.”

.....  
Ms. D. A. Meedeniya

.....  
Date



University of Moratuwa, Sri Lanka.  
Electronic Theses & Dissertations  
[www.lib.mrt.ac.lk](http://www.lib.mrt.ac.lk)

I certify that the declaration above by the candidate is true to the best of my knowledge and that this report is acceptable for evaluation for the M.Sc. in Computer Science.

.....  
Dr. A.S. Perera  
Senior lecturer,  
Department of Computer Science and Engineering,  
University of Moratuwa.

.....  
Date

To  
My Family


And



University of Moratuwa, Sri Lanka.  
Electronic Theses & Dissertations  
www.lib.mrt.ac.lk

## Abstract

The significant growth in the electronic media to store and exchange text documents has led to the use of tools, which analyse and categorize documents based on their content. The availability of full-text documents in electronic form emphasizes the need for intelligent information retrieval techniques. In Sri Lanka most of the public services use text documents written in the Sinhala language to provide their services. As a result, there is a need for systems that can be used to semi-automatically analyze and process documents in Sinhala. Wide availability of electronic data has led to the vast interest in text analysis, information retrieval and text categorization methods. There are many concepts, approaches and techniques associated with text mining. Most of the widely available text categorization tools work only with English text. Therefore to provide a better service, there is a need for non-English based document analysis and categorizing systems, as is currently available for English text documents.

 University of Moratuwa, Sri Lanka.  
A tool that can automatically categorize a collection of Sinhala documents can be an asset to any service provider that deals with a large number of text documents in Sinhala. Data clustering can be used to categorize documents based on the content. The effectiveness of clustering depends on the feature extraction. The main techniques examined in this study include data pre-processing, feature extraction, and document clustering. The approach makes use of a transformation based on the text frequency and the inverse document frequency, which enhances the clustering performance. This approach is based on Latent Semantic Analysis. A text corpus categorized by human readers is utilized to test the validity of the suggested approach.

The technique introduced in this work enables the processing of text documents written in Sinhala, and empowers citizens and organizations to do their daily work efficiently.

## **Acknowledgement**

It gives me great pleasure to acknowledge the contribution and assistance of a large number of individuals who made this research a complete success. Without their help and guidance I would not have been able to gain a successful research experience. Though it is hard to acknowledge everyone individually, there are some people who should be acknowledged, at least by saying a few words about them.

First of all I would like to thank my supervisor Dr. Shehan Perera, for accepting me as a research student and for the guidance he provided. Despite his busy work schedule, he was always supportive and guided me on the correct path. He was really helpful to me when I had doubts about the research activities. Specially, Dr. Perera's valuable effort in driving me to do an affluent research is highly appreciated.

My special thanks are extended to Prof. Gihan Dias, who was my first M.Sc. research co-ordinator, and Dr. Sanath Jayasena, the second M.Sc. research co-ordinator for my research. It was really a privilege to have such senior faculty members, who indeed, helped me to improve my work.

I would also thank to Mrs. Vishaka Nanayakkara, the Head of the Department of Computer Science and Engineering, for the support she gave to make this research a success. I would like to thank all the other lecturers and staff members of the Department of Computer Science and Engineering for the support they have given me in numerous ways. I would like to thank my colleagues who studied with me for the M.Sc. at the Department. It was a wonderful experience to be with them.

Finally, I am grateful to my family for encouraging me. They also helped me in whatever way they could, even helping me to cover up my daily chores when I was busy with this study.

In addition to the people mentioned above, if I have inadvertently not mentioned anyone, I should apologies about it. I appreciate everyone who helped me in my research.

D.A.Meedeniya

Dept. of Computer Science and Engineering, University of Moratuwa, Sri Lanka.

## Table of Content

Declaration .....	ii
Abstract .....	iv
Acknowledgement .....	v
List of Figures .....	ix
List of Tables .....	x
Chapter 1 – Introduction .....	1
1.1 Background .....	1
1.2 Problem Statement .....	3
1.3 Objective of the Study .....	4
1.4 Significance of the Study .....	5
1.5 Quality Requirements of the Study .....	7
1.6 Scope of the Study .....	7
1.7 Deliverables and Expected Outcomes .....	7
1.8 Resource Requirements .....	8
1.9 Structure of the Thesis .....	9
Chapter 2 - Literature Survey .....	10
2.1 Introduction .....	10
2.1.1 Data Mining .....	10
2.1.2 Text Mining .....	11
2.2 Text Categorization .....	12
2.2.1 Importance of Text Categorization .....	12
2.3 Text Categorizing Application Areas .....	12
2.3.1 Areas of Text Categorization .....	12
2.3.2 Applications of Text Categorization .....	13
2.4 Text Categorizing Approaches .....	14
2.4.1 Clustering .....	15
2.4.1.1. Unsupervised Clustering .....	17
2.5 Stages in Text Categorization .....	18
2.5.1 Data Pre-processing .....	18
2.5.2 Filtering Raw Data .....	19
2.5.3 Word Frequency .....	19
2.5.4 Transformations: .....	19
2.5.5 Information Retrieval .....	20
2.5.6 Feature Extraction .....	20
2.5.6.1. Language identification systems .....	20
2.5.6.2. Current Issues .....	20

2.5.6.3. Bag-of-words .....	21
2.5.6.4. Attribute selection.....	21
2.5.6.5. N-gram language identification .....	21
2.5.6.6. Trigrams identification.....	22
2.5.7 Vector Model .....	22
2.5.8 Dimensionality Reduction .....	23
2.6 Text Categorizing Learning Methods .....	24
2.6.1 Support Vector Machine .....	24
2.6.2 Self Organizing Map.....	25
2.6.3 Adaptive Resonance Associative Map (ARAM).....	25
2.6.4 Latent Semantic Indexing .....	26
2.6.5 Latent Dirichlet Allocation .....	27
2.6.6 K-Nearest Neighbour Algorithm .....	28
2.6.7 K-Means Clustering Algorithm .....	29
2.6.8 Gaussian Mixture Model (GMM).....	32
2.6.9 Principal Direction Divisive Portioning (PDDP) Clustering Algorithm .....	35
2.6.10 Comparison .....	35
2.6.11 Other Approaches Associated with Text Mining.....	36
2.7 Problems with Current Text Categorization .....	36
2.8 Text Categorizing Tools .....	37
2.8.1 Magenta Technology .....	38
2.8.2 WEBSOM Method and Browsing Interface .....	39
2.8.3 Other Tools for Text Mining.....	40
2.9 Sinhala Language.....	41
2.9.1 Features of Sinhala Language.....	41
2.10 Text Representation .....	42
2.10.1 ASCII Representation .....	43
2.10.2 ISO 8859 Representation .....	43
2.10.3 UNICODE Representation.....	43
2.11 Cluster Evaluation Techniques .....	44
2.11.1 Correlation .....	44
2.11.2 Checker-board Graph.....	45
2.11.3 Selecting the best partition.....	45
2.12 Performance .....	46
2.12.1 Recall and Precision.....	46
Chapter 3 - Case Study: Data Pre-Processing .....	48
3.1 Test Case based on Data Pre-Processing .....	48
3.2 Concept Model Case based on Agent Technology .....	51

Chapter 4 – Experiment Methodology.....	53
4.1 Methodology Overview .....	53
4.2 Implementation Procedure .....	54
4.2.1 Data Pre-Processing Stage .....	55
4.2.2 Pre-Text Categorizing Stage .....	56
4.2.3 Clustering Procedure.....	59
4.2.4 Analysis Process .....	63
4.2.5 Implementation Algorithm.....	63
Chapter 5 – Testing.....	65
5.1 Test Case Results: Based on data pre-processing .....	65
5.1.1 Obtained Results .....	66
5.1.2 Results Analysis.....	66
5.2 Experimental Methodology Test Results.....	67
5.2.1 Data Pre-Processing Test Results .....	67
5.3 Text Categorization Test Results .....	69
5.3.1 Results Analysis: Latent Semantic Analysis with SVD .....	69
5.3.2 Results Analysis: Clustering Techniques.....	72
5.4 Performance Evaluation.....	77
5.5 Concluding Remarks.....	80
Chapter 6 – Discussion .....	81
6.1 Study Outcome.....	81
6.2 Experimental limitations .....	84
6.3 Future Work.....	84
6.4 Summary .....	85
Conclusion .....	86
References.....	87
Appendix.....	93
Annex 1 .....	94



## List of Figures

Figure 2.1: Graphical representation of four clusters .....	16
Figure 2.2: Data points in a support vector machine .....	24
Figure 2.3: Comparison of data using checker-board graph.....	45
Figure 4.1: Major task of the methodology .....	53
Figure 4.2 Implementation Procedure.....	54
Figure 4.3 Sample of term x document frequency matrix .....	56
Figure 4.4 Steps of the GMM clustering .....	61
Figure 4.5 Steps of the k-means clustering.....	62
Figure 5.1 Content Representation .....	66
Figure 5.2 Title Representation.....	66
Figure 5.3 Sample of a data file .....	68
Figure 5.4 Sample of common words file.....	68
Figure 5.5 Sample of formatted data file .....	68
Figure 5.6 Sample of word frequency matrix .....	69
Figure 5.7 Sample of word frequency matrix .....	69
Figure 5.8 Graph of the Similarity matrix obtained by the correlation .....	70
Figure 5.9: Hierarchical representation of the cluster separations.....	70
Figure 5.10: Relationships among the sample documents.....	71
Figure 5.11: Checker-board graph of the coefficient of determination .....	72
Figure 5.12 Class mean of the Gaussian Mixture Model clustering.....	73
Figure 5.13: Checker-board graph of the identified clusters .....	73
Figure 5.14 Document grouping with two clusters.....	74
Figure 5.15 Probabilities of the documents in cluster 1.....	74
Figure 5.16 Probabilities of the documents in cluster 2.....	74
Figure 5.17 Document grouping with three clusters.....	75
Figure 5.18 Probabilities of the documents in cluster 1.....	75
Figure 5.19 Probabilities of the documents in cluster 2.....	75
Figure 5.20 Probabilities of the documents in cluster 3.....	76
Figure 5.21 Time taken for Document pre-processing .....	76
Figure 5.22 Time taken by different clustering techniques .....	77
Figure 5.23 Time taken by different clustering methods .....	77

Figure 5.24 Accuracy of the categorization .....	78
Figure 5.25 Percentage of the accuracy with different techniques .....	78

## List of Tables

Table 5.1 Results obtained on the document title data set.....	65
Table 5.2 Results obtained on the document content data set .....	66
Table 5.3 Analysis of Results .....	67
Table 5.4 Similarity matrix for the sample data. ....	69
Table 5.5 Coefficient of determination values between each documents.....	71
Table 5.6 Posterior values for each document after GMM clustering.....	73
Table 5.7 Document grouping with two clusters .....	74
Table 5.8 Probabilities of the documents in a given cluster .....	74
Table 5.9 Document grouping with three clusters .....	75
Table 5.10 Probabilities of the documents in a given cluster .....	75
Table 5.11 Comparison of clustering: Human vs. System.....	76



University of Moratuwa, Sri Lanka.  
 Electronic Theses & Dissertations  
[www.lib.mrt.ac.lk](http://www.lib.mrt.ac.lk)