Investigating the Applicability of Partition–Based Clustering for Sinhala Documents

D.A.Meedeniya

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

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 Date

Senior lecturer,

Department of Computer Science and Engineering,

University of Moratuwa.

To

My Family

And

Teachers

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.

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

List of Figures

List of Tables