Universiti Teknologi MARA

Indexing System for Spoken Document

Aishul Aman Bin Sulaiman Bukhari

Thesis submitted in fulfillment of the requirement for

Bachelor of Computer Science (Hons)

Faculty of Computer and Mathematical Sciences

MAY 2011

ACKNOWLEDGEMENT

In the name Allah, The All Merciful, The Most Merciful

Praise to Allah S.W.T. The Most Gracious.

I would like to most sincere acknowledgement and gratitude to my supervisor, Puan Haslizatul Fairuz Binti Mohamed Hanum for her guidance, advices and moral support for this project from the beginning towards the end. Her encouragement and ideas for this project were a source for my enthusiasm and inspiration.

I am also thankful to all my lecturers especially Dr. Riaza Perveen Binti Mohd Rias for her guidance on how the proposal should work and constantly supervised the project flow.

Last but not least, I would like to thank for all my friends who getting involved in this project. Their determination towards the works is really important in achieving the completion of the project. Another appreciation I would to express my thankful for my family. Their moral support gives me the strength to complete this project in time even though there are obstacles during the timeline. Again, thank you very much to all of you.

ABSTRACT

Retrieving relevant speeches including their respective video and audio from Hansard Document is one of the current problems in the parliament of Malaysia. The website of the parliament of Malaysia also only provides the text-based result and output in the system. In this project, the scopes are the spoken document from the parliament of Malaysia and randomly chosen audio and video files as the corresponding information for the spoken document. Next, this project discussed about the information retrieval and the indexing system applied to get the text also with its corresponding audio and video files before we can search the information using the Hansard Document precisely. The use of an algorithm of naming convention will be developed to create an indexing name for each document to make an ease in searching the requested information. Therefore, with a proper storage for the information and indexing method, we can get the relevant information more precisely and easily in a faster time. For the future enhancement, this system needs a good database for more accurately to store the raw data for retrieving information.

Keywords: Information Retrieval, indexing, naming convention

TABLE OF CONTENTS

SUPERVISOR APPROVAL	**ii**
DECLARATION	**iii**
ACKNOWLEDGEMENT	**iv**
ABSTRACT	**v**
TABLE OF CONTENTS	**vi**
LIST OF FIGURES	**vii**
LIST OF TABLES	**viii**

1. **Chapter 1**–Introduction 1

    1.1 Background 2
    1.2 Hansard Document 2
    1.3 Problem Statement 2
    1.4 Objective of the project 2
    1.5 Scope of the project 3
    1.6 Project aim 3
    1.7 Summary 3

2. **Chapter 2**– Literature Review 4

    2.1 Introduction 4
    2.2 Information Retrieval 4
    2.3 The Information Retrieval Process 5
    2.4 Indexing and Query Evaluation 6
        2.4.1 Index Construction 8
        2.4.2 Automatic Indexing 8

CHAPTER 1

INTRODUCTION

1.1 BACKGROUND

Methods of searching for information in documents, searching for documents themselves, searching for metadata which describe documents, or searching within database or databases such as the World Wide Web is called Information Retrieval (IR) (Mohammed Chaoui and Mohammed Tayeb Laskri, 2011).

Malaysia is a country that practices the style of administrative in the form of democracy. One of the administrations is the Parliament of Malaysia. Each year, the Parliament will have a session to meet up all the speakers in order to discussing the problems regarding on what have been happened in Malaysia. This session will be then recorded and documented to save all the discussions made by the speakers. The types of the documents and records are in PDF format for the text documentations and WMV format for the audio/video parts.

The recorded version of each session will be then updated and uploaded to the parliament's website. However, the record only contains text documents which mean there is only PDF file that already been uploaded to the website database. Any users or speakers that want to see what actually happened in the each session from parliament are limited. This is because only PDF or text documents can be accessed and downloaded from the parliament's website.