

UNIVERSITI TEKNOLOGI MARA

**ENHANCING LATENT SEMANTIC
ANALYSIS (LSA) USING TAGGING
ALGORITHM IN RETRIEVING
MALAY DOCUMENTS**

AFIQAH BAZLLA BINTI MD SOOM

Thesis submitted in fulfillment
of the requirements for the degree of
Master of Science
(Computer Science)

Faculty of Computer Science and Mathematics

March 2018

CONFIRMATION BY PANEL OF EXAMINERS

I certify that a Panel of Examiners has met on 18th September 2017 to conduct the final examination of Afiqah Bazlla binti Md Soom on his Master of Science thesis entitled “Enhancing Latent Semantic Analysis (LSA) using Tagging Algorithm in Retrieving Malay Documents” in accordance with Universiti Teknologi MARA Act 1976 (Akta 173). The Panel of Examiners recommends that the student be awarded the relevant degree. The Panel of Examiners was as follows:

Maheran Jaafar, PhD
Associate Professor
Faculty of Computer & Mathematical Sciences
Universiti Teknologi MARA
(Chairman)

Nasiroh Omar, PhD
Senior Lecturer
Faculty of Computer & Mathematical Sciences
Universiti Teknologi MARA
(Internal Examiner)

Saidah Saad, PhD
Senior Lecturer
Faculty of Information Science and Technology
Universiti Kebangsaan Malaysia
(External Examiner)


**PROF SR DR HJ ABDUL HADI
HJ NAWAWI**
Dean
Institute of Graduates Studies
Universiti Teknologi MARA
Date : 13 March 2018

AUTHOR'S DECLARATION

I declare that the work in this thesis was carried out in accordance with the regulations of Universiti Teknologi MARA. It is original and is the results of my own work, unless otherwise indicated or acknowledged as referenced work. This thesis has not been submitted to any other academic institution or non-academic institution for any degree or qualification.

I, hereby, acknowledge that I have been supplied with the Academic Rules and Regulations for Post Graduate, Universiti Teknologi MARA, regulating the conduct of my study and research.

Student I.D. No. : 2014634666
Programme : Master of Science (Computer Science)
CS750
Faculty : Computer Science and Mathematics
Thesis Title : Enhancing Latent Semantic Analysis (LSA)
using Tagging Algorithm in Retrieving Malay
Documents

Signature of Student : 

Date : March 2018

ABSTRACT

Latent Semantic Analysis (LSA) algorithm is a mathematical approach that uses Singular Value Decomposition (SVD) to discover the important association of the relationship between terms and terms, terms and documents and also documents and documents. Furthermore, LSA uses cosine similarity measurement to measure the similarity between the query word and terms as well as the documents. This approach seem to be efficient if each of the term only have single meaning and a meaning only represent a single term. Unfortunately, in Malay language there exists many terms that have multiple meanings and a single meaning that are represented by multiple terms. If these terms are treated as a single word, it will lead the search engine to retrieve irrelevant documents. These irrelevant documents retrieved will effect the effectiveness of the search engine. To investigate the enhancement of LSA using tagging algorithm (LSAT) in retrieving Malay documents, eight experiments are conducted in this research. The first experiment is conducted to compare the time taken for extracting normal term list and tagged term list, total number of both lists and also the time taken for the creation of term document matrix. Another six experiments record all the results of the LSA and LSAT search engine by using different dimension and threshold value. While the last experiment to compare the LSAT result with previous work on LSA using the same test collection. Outcomes of this study indicate that by using tagging algorithm, the recall value of the LSA algorithm can be enhanced up to 4% , the precision value also can be enhanced up to 16% and the F-measure value of LSA retrieval result can be enhanced by approximately up to 7% compared to LSA retrieval result without tagging algorithm. Furthermore, this research provides fundamental analyses to the other Information Retrieval (IR) developer in selecting the value of dimension and threshold value of retrieval that using LSA.

TABLE OF CONTENTS

	Page
CONFIRMATION BY FINAL OF EXAMINERS	ii
AUTHOR'S DECLARATION	iii
ABSTRACT	iv
ACKNOWLEDGEMENT	v
TABLE OF CONTENTS	vi
LIST OF TABLES	x
LIST OF FIGURES	xi
LIST OF SYMBOLS	xv
LIST OF ABBREVIATIONS	xvi
CHAPTER ONE: INTRODUCTION	1
1.1 Introduction	1
1.2 Project Background	2
1.3 Problem Statement	3
1.4 Research Objectives	4
1.5 Research Questions	6
1.6 Research Scopes	7
1.7 Significance Of The Research	7
1.8 Summary	8
CHAPTER TWO: LITERATURE REVIEW	9
2.1 Introduction	9
2.2 Information Retrieval (IR)	9
2.2.1 Basic Processes of Information Retrieval	10
2.2.2 Mathematical Models of Information Retrieval	11
2.2.3 Information Retrieval Modelling	11
2.2.4 Three Problems that Models of IR have to Solve	12