

**COMPARISON OF RETRIEVAL SCHEME BASED ON  
DIFFERENT STRUCTURE AND SIMILARITY MEASURE  
FOR MEMOS AT TATI UNIVERSITY COLLEGE –  
CENTRE OF HND STUDIES**

**BY**

**AINI ZURIYATI BT ABDUL KADIR**

**Project Paper Submitted in Partial Fulfillment of the Requirement for the Degree of  
Master of Information Technology**

**OPEN UNIVERSITY MALAYSIA**

Digital Library OUM



0030437

## ABSTRACT

Retrieval models and techniques can be applied to retrieve memo information and it does relate to certain queries or concepts. Different method of retrieval shows different relevant documents. Therefore, different methods are used to retrieve, structure and similarity measures being used to retrieve memo. The data are collected at Center of HND Studies at TATI University College. The process that has been applied is digitizing, stop word removal, stemming and building index. The results of this process then stored in database. In this study 50 document of memo being used, where refer to the titles and the document. In order to show the measure of different memo structure three similarity measures being used. The structure refers to the titles of memo, the documents and combination of both title and documents. The results shows that title structure have the lowest performance of result compared to document structure and the combination of both. Meanwhile, in similarity measures the results shows that weighting schemes using cosine perform well compared to Dice and Russell-Rao. The overall result shows that the best performance for retrieving memo is by using document structure coupled with Cosine similarity.

## TABLE OF CONTENTS

DEDICATION	iii
ABSTRACTS	iv
ABSTRAK	v
ACKNOWLEDGEMENTS	vi
APPROVAL	vii
DECLARATION	viii
TABLE OF CONTENTS	ix
LIST OF TABLES	xiii
LIST OF FIGURES	xiv
LIST OF SYMBOLS	xv
LIST OF ABBREVIATIONS	xiv

### Chapter 1

#### INTRODUCTION

1.1	Introduction	1
1.2	Problem Background	3
1.3	Problem Statement	4
1.4	Project Objectives	5
1.5	Project Scope	5
1.6	Significance of the Project	6
1.7	Organization of the Report	7

### Chapter 2

#### LITERATURE REVIEW

2.1	Introduction	8
2.2	Basic IR System	9
2.2.1	Stemming Algorithm	9

2.2.2	Nice Stemmer Algorithm	10
2.2.3	Text Stemmer Algorithm	11
2.2.4	Porter Stemmer Algorithm	11
2.2.5	Stop Words	11
2.3	Model of Information Retrieval	12
2.3.1	Boolean Model	12
2.3.2	Vector Space Model	14
2.3.3	Probabilistic Model	16
2.4	Data Model	18
2.4.1	Structure Data	18
2.4.2	Semi structured Data	19
2.5	Indexing	19
2.5.1	Block Merge Indexing	20
2.5.2	Distributed Indexing	21
2.5.3	Dynamic Indexing	22
2.6	Evaluation in Information Retrieval	22
2.6.1	Standard Benchmark	23
2.6.2	Retrieval Performance	25

**Chapter 3****METHODOLOGY**

3.1	Introduction	28
3.2.1	Operational Framework	28
3.2.1	Data model	29
3.2.2	Index Construction	30
3.2.3	Indexing	33
3.2.4	Memos Searching Technique	37
3.4	Similarity Measures	40
3.5	Instrumentation	41
3.6	Writing Report	42
3.7	Summary	42

**Chapter 4****DATA ANALYSIS AND RESULT**

4.1	Introduction	43
4.2	The Result of Cosine Similarity	43
4.3	Result of Dice Similarity	45
4.4	Result of Russell-Rao Similarity	46
4.5	Comparison of Similarity Measure	47
4.6	Conclusion of results	48

**Chapter 5****DISCUSSION**

5.1	Introduction	49
5.2	Experimental	49
5.3	Contribution of Study	50
5.4	Suggestion	50

**Chapter 6****SUMMARY AND CONCLUSION**

6.1	Summary	51
6.2	Conclusion	52

**References**

53

**Appendices**

55

Appendix A: Algorithm Used In the Study

56

Appendix B: List of Stop Word

58

Appendix C: Porter Stemmer Algorithm

62

Appendix D: List of Query

68

Appendix E: Inverted File Structure

70

Appendix F: Example of Memo

72

## LIST OF TABLES

TABLE NO.	TITLE	PAGE
2.1	Advantages and disadvantages of Standard Boolean Model.	13
2.2	List of key advantages and disadvantages of Vector Space Model.	15
2.3	List of key advantages and disadvantages of Probabilistic Model.	17
3.1	The example of stop words	32
3.2	Indexing (part one).	35
3.3	Indexing (part two)	36
4.1	The performance of individual weighting using Cosine measure	44
4.2	The performance of individual weighting using Dice measure.	45
4.3	The performance of individual weighting using Russel-Rao measure	46
4.4	Shows the comparison of the measure between Cosine, Dice and Russell-Rao	47

## LIST OF FIGURES

FIGURE NO.	TITLE	PAGE
2.1	General components of the information retrieval process	9
2.2	The mathematical basis and the properties of the IR model	12
3.1	Components of vector data model and retrieval technique	28
3.2	Data model for retrieving system	29
3.3	Steps in index constructions	30
3.4	Main Searching Techniques	37
4.1	Comparison of indexing structure using Cosine measure	44
4.2	Comparison of indexing structure using Dice measure	45
4.3	Comparison of all structure indexing Russell-Rao measure	46
4.4	Comparison between the measure of Cosine, Dice and Russell-Rao	47



## Chapter 1

### INTRODUCTION

#### 1.1 Introduction

Information retrieval (IR) is finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need from within large collections (usually stored on computers). As defined in this way, information retrieval used to be an activity that only a few people engaged in: reference librarians, paralegals, and similar professional searchers. Now the world has changed, and hundreds of millions of people engage in information retrieval every day when they use a web search engine or search their email.

The field of information retrieval also covers operations typically done in browsing document collections or further processing a set of retrieved documents. Given a set of documents, clustering is the task of coming up with a good grouping of the documents based on their contents. It is similar to arranging books on a bookshelf according to their topic. Given a set of topics (or standing information needs), classification is the task of deciding which topic(s), if any, each of a set of documents belongs to. It is often approached by first manually