COMPARISON AND FUSION OF RETRIEVAL SCHEMES BASED ON DIFFERENT STRUCTURES, SIMILARITY MEASURES AND WEIGHTING SCHEMES

MOHAMMED SALEM FARAG BIN WAHLAN

UNIVERSITI TEKNOLOGI MALAYSIA

COMPARISON AND FUSION OF RETRIEVAL SCHEMES BASED ON DIFFERENT STRUCTURES, SIMILARITY MEASURES AND WEIGHTING SCHEMES

MOHAMMED SALEM FARAG BIN WAHLAN

A project report submitted in partial fulfillment of the
requirements for the award of the degree of
Master of Science (Computer Science)

Faculty of Computer Science and Information System
Universiti Teknologi Malaysia

MARCH 2006

DEDICATION

This thesis is dedicated to my beloved family and to whoever serves the truth for the truth itself.

ACKNOWLEDGEMENTS

First and foremost, I would like to thank ALLAH S.W.T. for all the achievements that I have gained today. Next, I wish to extend my grateful appreciation to all those who have contributed directly and indirectly to the preparation of this study. I would like to take this opportunity to thank my supervisor, Associate Professor Dr. Naomie Salim for attention, encouragement and guidance throughout the length of this study. Not forgetting my beloved family for all the supports and understandings that they have given to me. Not forgetting also, my examiners Associate Professor Dr. Mohd Nor bin Mohd Sap and Dr. Rubaiah Ahmed for many helpful suggestions.

I am grateful to all my colleagues, friends, staff, and lecturers in Faculty of Computer Science and Information System, Universiti Teknologi Malaysia and Hadhramout University of Science and Technology for their help and support at every step during this course of studies.

ABSTRACT

Many retrieval models and techniques can be applied to retrieve theses that are most relevant to certain queries or concepts. It has been found that different retrieval methods often retrieve different sets of relevant documents. It is therefore anticipated that a particular retrieval method will usually retrieve some relevant theses not retrieved by other methods. Therefore in this study, different methods are used in the theses retrieval, based on different thesis structures, different similarity measures and different weighting schemes. The theses used in this study are collected from FSKSM postgraduate library. Many operations have been applied on the collected theses such as digitizing, stop words removal, stemming and building index. The results from these operations are stored in a database. In this study, 85 theses and 30 queries are used. The comparisons between query and theses were made using five similarity measures with seven weighting schemes using different thesis structures. The results show that the use of bibliography gives poorer results compared to the use of title and abstract alone. In the weighting schemes combinations, the results show that weighting schemes using Cosine and Tanimoto perform well individually but did not do well in the combinations and weighting schemes using Forbes and Russell similarity measures do not do well individually but did well in the combination. In the similarity measures combinations, the results show that the best combination was Cosine using LTU weighting scheme with Russell using LOGG weighting scheme using title structure but using abstract structure, the best combination was Cosine using TFIDF weighting scheme with Forbes using ATFA weighting scheme but it has less performance than the combination of Cosine using LTU weighting scheme with Russell using LOGG weighting scheme using title structure. The overall results show that the best thesis structure is title and the best similarity measure is Cosine with LTU weighting scheme.

ABSTRAK

Terdapat banyak model dan teknik pengembalian maklumat yang telah diaplikasikan dalam pelbagai domain kajian. Hasil set dokumen berbeza jika kaedah pengembalian maklumat berbeza. Kaedah yang digunakan dalam domain kajian ialah pengembalian tesis. Struktur tesis yang berlainan juga skema pemberat yang berbeza akan diaplikasikan dalam kaedah-kaedah yang digunakan. Sebanyak 85 tesis yang diperolehi daripada Pejabat Pasca Ijazah FSKSM telah digunakan untuk 30 queries. Prapemprosesan yang terlibat ke atas tesis-tesis ini termasuklah pendigitalan, penghapusan perkataan-henti, pembuangan akar perkataan serta pembinaan indeks. Selanjutnya, hasil-hasil prapemprosesan ini disimpan di dalam pangkalan data. Perbandingan di antara query dan tesis dilaksanakan berdasarkan kepada lima ukuran persamaan beserta tujuh skema pemberat di mana struktur tesis yang berlainan akan digunakan. Penggunaan bibliografi menunjukkan hasil yang kurang memuaskan berbanding penggunaan tajuk atau abstrak. Secara individu, Cosine dan Tanimoto memberikan keputusan yang memuaskan untuk kombinasi skema pemberat dan sebaliknya hasil keputusan kurang memuaskan dalam kombinasi dan skema pemberat menggunakan persamaan Forbes. Manakala persamaan Russell memberikan hasil yang kurang memuaskan secara individu berbanding hasil keputusan kombinasi. Dalam pada itu, gabungan ukuran persamaan menunjukkan gabungan Cosine menggunakan skema pemberat LTU dengan Rusell menggunakan skema pemberat LOGG memberikan hasil yang terbaik bagi struktur tajuk. Manakala bagi penggunaan struktur abstrak, gabungan terbaik adalah daripada gabungan Cosine menggunakan skema pemberat TFIDF dengan Forbes menggunakan skema pemberat ATFA. Keseluruhan keputusan menunjukkan struktur tajuk merupakan struktur yang terbaik bagi struktur tesis manakala Cosine dengan skema pemberat LTU merupakan ukuran persamaan yang terbaik dalam kajian ini.

TABLE OF CONTENT

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS

R	Set of Relevant Document
|R|	Number of Documents in the Set R
A	A document Answer Set
|A|	Number of Document in the Set A
|Ra|	Number of Documents in the Intersection of the Sets R and A
D.q'	Dot Product Between Document Terms and Query Terms
∑	Sum
√	Square Root

LIST OF ABBREVIATION

IR	Information Retrieval
IDF	Inverse Document Frequency
TF	Term Frequency
CF	Concept Frequency
NORM	Document Normalization Factor
NTF	Normalized Term Frequency Factor
TREC	Text Retrieval Conference
NLTF	Non-logarithmic Term Frequency
P_R	Precision and Recall
XML	Extensible Markup Language

LIST OF APPENDICES

CHAPTER 1

INTRODUCTION

1.1 Introduction

The aim of information retrieval is to provide the user with the “best possible" information from a database. The problem of information retrieval is determining what constitutes the best possible information for a given user query. A common form of interaction for information retrieval is for the user query. These are then used by the information retrieval system to identify information that meets the user’s needs. For example, in a bibliographic database, a user might be interested in finding thesis on some topic. The keywords extracted from the query would be an attempt to delineate that topic, and then used to improve precision (ensuring that a significant proportion of the items retrieved are relevant to the user) and recall (ensuring that a significant proportion of the relevant items are retrieved).

Modern IR systems accept free-format natural language queries from users. A query is said to represent the “information need” of the user. Given a large collection of documents, a small subset containing one or more key words from the query statement is retrieved by the IR system. The IR system usually employs some method to “predict” the relevance of a document. Documents retrieved are ranked in decreasing order of their predicted relevance (Wensi, 2000).

Given a user query, a good information retrieval system would rank most of the relevant documents ahead of less relevant documents, thereby allowing the user

to peruse relevant documents without having to wade through many irrelevant documents.

Several retrieval models and techniques have been developed for information retrieval (Frakes and Baeza, 1992). It has been found that different retrieval methods often retrieve different sets of relevant documents. A particular retrieval method will usually retrieve some relevant documents not retrieved by other methods. In this thesis, we will explore thesis retrieval based on different structures (title, abstract and bibliography), weighting schemes and similarity measures. We will also study whether data fusion from different retrieval approaches can give better results compared to singular approach.

1.2 Problem Background

The information retrieval presupposes that there are some documents or records containing information that have been organized in an order suitable for easy retrieval (Chowdhury, 1999). The main problem in achieving an efficient and user-friendly retrieval is the development of a search mechanism to guarantee delivery of minimal irrelevant information (high precision) while insuring relevant information is not overlooked (high recall).

The performance of retrieval process is affected by many factors like weighting schemes, similarity measures, retrieval models, and document structures. Many studies have been done to achieve the best performance and discover the factors which affect the retrieval. For instance, Liu and Croft, (2002) have compared the passage retrieval with full text and found that passage can provide more reliable performance than full text. Park, *et. al*, (2003) have compared the title of web page with other page sections using *tf* weighting scheme and found that giving more importance to title section in web page leads to performance improvement.

The weighting schemes in the performance of information retrieval systems are important factors. There are many studies for evaluating the performance of different term weighting schemes, Jin, *et. al*, (2001) have compared four term weighting schemes "nnn", "atc", "ltu" and "Okapi" and found that "ltu " and "okapi" are the best term weighting schemes. (Hersh, 1994) has compared many term weighting schemes and found that the best single term weighting scheme was the inverse document frequency (IDF) and the best performance for weighting formula occurred with the combination of factors (IDFi*TFij)/(CFij*NORMj).

A similarity measure is any function which assigns a number to a pair of vectors. Simple similarity measures may count number of terms in agreement between query and document. This study uses five similarity measures; four of those similarity measures (Cosine, Russell, Forbes and Tanimoto) use the weighting scheme after calculating it. Salim (2002) has found that those similarity measures (Cosine, Russell, Forbes and Tanimoto) are the best and perform well. The fifth similarity measure called Okapi has formula in which the weighting scheme calculated directly. This measure has been evaluated thoroughly in the context of NIST's TREC information retrieval and has been found to be especially powerful (Kemp and Waibet, 1998). For this reason, those similarity measures will be used in this study.

The combination of different text representations and search strategies has become a standard technique for improving the effectiveness of information retrieval. Many works on data fusion have been done to improve the retrieval result, (Wensi, 2000) has combined the traditional tf*idf weighting scheme with a weighted binary weighting scheme and found the retrieval effectiveness improved by 53% for long queries and 90% for short queries. Lee, (1995) has combined two retrieval runs in which one performs cosine normalization and the other does not, he found significant improvements. Tzitzikas, (2001) has combined results from different system and proposed a fusion technique. The choosing of a good combination of coefficients can lead to the best use of fusion (Salim, 2002). One of the reasons to use data fusion in information retrieval, is that no system can give an interpretation about a topic that can completely capture a unique meaning for all readers.

1.3 Problem Statement

This project aims to provide a comprehensive comparison of search schemes based on different structures like search by title, abstract and bibliography for finding out which one is better than others in finding relevant theses to a user query. In addition, different similarity measures and different weighting schemes will also be used on the different structures and a comparison will be made on which structure used with which weighting scheme is better than others for retrieving thesis most similar to a new project/thesis based on its title. The fusion of weighting schemes and similarity measures will be made to explore whether it can improve the performance. The main problem in achieving an efficient and user-friendly retrieval is the development of a search mechanism to guarantee delivery of minimal irrelevant information (high precision) while insuring relevant information is not overlooked (high recall).

There are three issues that can achieve the goal of this study: *what is the thesis structure which provides better performance in retrieval process?, what is the weighting scheme and similarity measure that can be used together with the structure that can better retrieve similar thesis to a title at hand? And what is the best combination of weighting schemes and similarity measures which can have a good performance?.*

1.4 Project Aims and Objectives:

The study aims to investigate thesis retrieval process using different structures, similarity measures and weighting schemes and achieving a search mechanism to guarantee delivery of minimal irrelevant information (high precision) while insuring relevant information is not overlooked (high recall) for effective retrieval.

Specific objectives of this project are:

1. To analyze performance of the thesis retrieval based on title, abstract and bibliography structures to understand which thesis structure can improve the performance.
2. To analyze performance of thesis retrieval based on different weighting schemes, similarity measures and their fusion on the different structures in (1) above to understand which weighting scheme, similarity measure and their fusion performs well.

1.5 Project Scope:

1. Title, abstract and bibliography of 85 FSKSM postgraduate theses will be stored on the machine and will be used in information retrieval process.
2. This project will make use of three thesis structures, seven weighting schemes (TFIDF, NORM, ATFA, LOGG, IGFI,IGFS and LTU), five similarity measures (Cosine, Russell, Forbes and Tanimoto) and their fusion.
3. This project will focus on Vector Space Model.
4. Only Porter stemming algorithm and inverted index will be used.
5. Relevance will be based on two expert's evaluation.
6. Data fusion is based on simple summation technique.

1.6 Significance of the Project

This study gives insight on what weighting scheme, similarity measure, structure and combination of them works best to retrieve thesis structure.

1.7 Organization of Report

Chapter 2 discusses the literature review. Chapter 3 discusses on the methodology used to build up this project. Chapter 4 discusses results. Chapter 5 presents the conclusion of this study.

REFERENCES

Aalbersberg I.J. (1994). *A Document Retrieval Model Based on Term Frequency* Ranks. Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval.1994. New York: Springer-Verlag.

Aljlayl, M. Frieder, O. and Grossman, D. (2002). *On Arabic-English Cross-Language Information Retrieval: A Machine Translation Approach.* International Conference on Information Technology: Coding and Computing (ITCC). April, 2002. Nevada: Las Vegas.

Asian, J. Williams, H.E. and Tahaghoghi S.M.M. (2004). *A Testbed for Indonesian Text Retrieval.* In *Proceedings of the 9th Australasian Document Computing Symposium*. December *2004*. Melbourne, Australia, 55-58.

Baeza-Yates, R., & Ribeiro-Neto, B. (1999), *Modern Information Retrieval*, ACM Press 1999.

Bear, J., Petit, J., and Martin, D. (1998). *Using Information Extraction to Improve Document Retrieval*, TREC-6, 367-378.

Chowdhury, G. G. (1999). *Introduction to modern information retrieval*. London: Library Association Publishing.

Dhillon, I. .S., Fan, J., and Guan, Y. (2001). *Efficient clustering of very large document collections*. In V. K. R. Grossman, C. Kamath and R. Namburu, editors, Data Mining for Scientific and Engineering Applications. Kluwer Academic Publishers.

Frakes, W. and Baeza-Yates, R. *(*1992). *Information Retrieval: Data Structure and Algorithms.* Englewood Cliffs, NJ.:Prentice Hall.

Fuller, M.l and Zobel J. (1998). *Conflation-based Comparison of Stemming Algorithms.*, In Proceedings of the Third Australian Document Computing Symposium Sydney, Australia, 21st August, 1998.

Greiff, W.R., Morgan, W.T and. Ponte, J.M. (2002). *Information retrieval models: The role of variance in term weighting for probabilistic information retrieval.* Proceedings of CIKM 2002.

Harman, D.K. (1993). *Overview of the first Text Retrieval Conference (TREC-1). In Proceedings of the First Text Retrieval Conference (TREC-1),* 1-20. NIST Special Publication, 500-207.

Harmanani, H.M., Keirouz, W.T. and Raheel, S. (2004). A *Rule-Based Extensible Stemmer for Information Retrieval With Application To Arabic.* Proceedings of the 8th IASTED International Conference on Artificial Intelligence and Soft Computing. September 2004. Spain, 35-40.

Hersh, W.R., Hickam, D.H., Haynes, R.B., and McKibbon, K.A. (1994). *A performance And Failure Analysis of SAPHIRE with a MEADLINE Test Collection.* Journal of the American Medical Informatics Association. 1(1): 51-60.

Hull, D. (1993). Using *Statistical Testing in the Evaluation of Retrieval Experiments*. In Proceedings of the Sixteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'93). 1993. USA: ACM Press, 329-338.

Jin, R., Falusos, C. and Hauptmann, A.G.(2001). *Meta-scoring: Automatically Evaluating Term Weighting Schemes In IR Without Precision-Recall*. In Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval. 2001. United States, 83-89.

Katzer, J., McGill, M.J., Tessier, J.A., Frakes, W. and Dasgupta, P. (1982). *A study of the overlap among document representations*. ACM Press.17(4): 106–114.

Kemp, T. and A. Waibel (1998). *Reducing the OOV rate in broadcast news speech recognition*. In Proceedings of the International Conference on Spoken Language Processing, 1839-1842.

Khan, L.R. *(1999). Ontology-based Information Selection.* University Of Southern California: Ph.D. Thesis.

Lee, J.H. (1995). *Combining Multiple Evidence From Different Properties Of Weighting Schemes.* In E. A. Fox, P. Ingwersen, and R. Fidel, editors, Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. 1995. USA: ACM Press, 180-188.

Liu, X. and Croft, W.B.(2002). *Passage Retrieval Based On Language Models.* In Proceedings of the eleventh international conference on Information and knowledge management. 2002. ACM Press: 375-382.

Magnini, B. and Prevete, R.(2000). *Exploiting Lexical Expansions and Boolean Compositions for Web Querying.* ACL'2000 Workshop on Recent Advances in Natural Language Processing and Information Retrieval. 8 October 2000. Hong-Kong University of Science and Technology.

Nascimento, M.A. and Cunha, A.C.R., An Experiment Stemming Non-Traditional Text (with Cunha). SPIRE'98, Proceedings, 75-80.

Park, E., Ra Dong-Yul and Jang, M. (2003). *Techniques for improving web retrieval effectiveness.* Information Processing & Management. 41(5): 1207-1223.

Porter, M.F. (1980). An Algorithm for Suffix Stripping. *Program – Automated Library and Information Systems*, 14(3): 130-137.

Rijsbergen, C.J. (1979). *Information Retrieval*. 2nd Edition.London: Butterworths.

Roberts, I. and Gaizauskas, R. *(*2004*). Evaluating Passage Retrieval Approaches for Question Answering.* In Proceedins of the 26th European Conference on Information Retrieval Research (ECIR 2004), Lecture Notes in Computer Science 2997. Heidelberg: Springer-Verlag, 72-84.

Salim, N. (2002). *Analysis and Comparison of Molecular Similarity Measures*. University of Sheffield: Ph. D. Thesis.

Smeaton, A.F. (1998). *Independence of Contributing Retrieval Strategies in Data Fusion for Effective Information Retrieval*. In Proceedings of the 20th BCS-IRSG Colloquium, Grenoble. April 1998. France, Springer-Verlag Workshops in Computing.

Teevan, J.B. (2001). *Improving Information Retrieval with Textual Analysis: Bayesian Models and Beyond.* Massachusetts Institute Of Technology: Master thesis.

Terra, E. and Clarke, C.L.A.(2004). *Information retrieval models: Scoring missing terms in information retrieval tasks* In *Proceedings of the 13th ACM Conference on Information and Knowledge Management* (CIKM). November 2004. Washington DC, USA.

Tzitzikas, Y. (2001). *Democratic Data Fusion for Information Retrieval Mediators.* In ACS/IEEE International Conference on Computer Systems and Applications,. June 2001. Beirut, Lebanon.

WENSI, X.(2000). *Combining Multiple Source of Evidence for Information.* Nanyang Technological University: Master Thesis.

Wilkinson, R., Zobel, J. and Sacks, D.R. (1995) *Similarity Measures for Short Queries.* . In D.K. Harman, editor, Proceedings of the 4th Text Retrieval Conference TREC-4. 1995. 277-286. NIST Special Publication 500-236.

Xu, J. Fraser, A. and Weischedel, R.(2002). *Empirical Studies in Strategies for Arabic Retrieval.* In *SIGIR 2002*, Tampere, Finland: ACM, 2002.

YU, C. T. (1982). *Term Weighting in Information Retrieval Using the Term Precision Model*. ACM Press. 29(1): 152-170.