



Faculty of Information Technology

**AUTOMATED TEXT CLASSIFIER OF ELECTRONIC
THESIS AND DISSERTATION**

CHAN AI LING

UNIVERSITI MALAYSIA SARAWAK

2003

QA
76.76
C454
2003

AUTOMATED TEXT CLASSIFIER OF ELECTRONIC THESIS AND DISSERTATIONS

**P.KHIDMAT MAKLUMAT AKADEMIK
UNIMAS**



0000112302

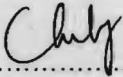
Chan Ai Ling

**A report submitted
in partial fulfillment of the requirements for the degree of
Bachelor of Information Technology**

**Faculty of Information Technology
UNIVERSITI MALAYSIA SARAWAK
March 2003**

DECLARATION

No portion of the work referred to in this report has been submitted in support of an application for another degree or qualification of this or any other university or institution of higher learning.



.....
(Signature)

8/3/2003

.....
(Date)

QA
76.76
C454

ACKNOWLEDGEMENT

I would like to thank the following people whose support throughout this project.

First of all, I would like to dedicate my highest appreciation to my supervisor, Prof. Madya K. Narayanan and Mr. Bong Chih How. I would like to thank them in advising and guiding me throughout the time in building this project and prototype.

Secondly, I would like to thank Ms. Suit Wai Yeng, my ex-supervisor in my ex-training company. She helps me a lot in solving the technical term in processing the prototype.

I would like to express my gratitude to my beloved, Liew Chih Hua who has helped me in this research in success.

Finally, I would like to dedicate my special thanks to my parent and family, friends and course-mates. Thanks for their fully supports and understanding.

Table of Contents

Declaration	ii
Acknowledgements	iii
Table of contents	iv
List of Tables	vii
List of Figures	viii
Abstract	ix
Abstrak	x
1. Chapter 1 : Introduction	1
1.1 Introduction	1
1.2 Problem statement	2
1.2.1 Problem solution	2
1.3 Automated text classifier for Electronic Thesis and Dissertations system	4
1.4 Objectives	4
1.5 Scope of project	5
1.6 Research significance	6
1.7 Project Plan	6
1.8 Outline of Project Report	7
2. Chapter 2 : Literature Review	8
2.1 Introduction	8
2.2 History	8
2.2.1 Electronic Thesis and Dissertation system	8
2.3 Reviewing of Existing system	8
2.3.1 Reviewing of Electronic and Dissertation System	9
2.3.1.1 Virginia Tech's Electronic Thesis and Dissertations System	9
2.3.1.2 Networked Digital Library of Theses and Dissertations	9
2.3.2 Comparison of the functions	9
2.4 System Architecture	11
2.4.1 Information Retrieval System	11
2.4.1.1 Broad outline of an Information Retrieval System	12
2.4.1.2 Retrieval Modes	12
2.4.1.2.1 Boolean search model	13
2.4.1.2.2 Vector Processing model	13
2.4.1.2.3 Best Match Searching Model	13
2.4.1.3 Evaluation of an Information Retrieval System	13
2.5 Proposed System	14
2.5.1 Improvements of current system	14
2.6 Comparison of Implementing Tools	14
2.6.1 Automated Text Classifier	14
2.7 Conclusion	15
3. Chapter 3 : Methodology	16
3.1 Introduction	16
3.2 Methodology	16
3.2.1 Problems and Objectives Identification	16
3.2.2 Literature Review	16

3.2.3 Identify System Needs	18
3.2.4 System Design	18
3.2.5 System Implementation	18
3.2.6 Experimentation	18
3.2.7 Result and Analysis	18
3.3 Summary	18
4. Chapter 4 : System Design	19
4.1 Introduction	19
4.2 Automated text classifier framework	19
4.2.1 Extract Portable Document Format (PDF) Information	21
4.2.2 Stemming	21
4.2.3 Index Terms List Matching	21
4.2.4 Document Vector	22
4.2.5 Similarity Measurement	22
4.2.6 Similarity Metric	24
4.2.7 Classification	24
4.3 Requirement specification	24
4.3.1 System requirement	24
4.3.1.1 Relevant electronic thesis and dissertations retrieval	24
4.3.1.2 New electronic thesis and dissertations classification	24
4.3.1.3 Search result accuracy	25
4.3.2 User requirement	25
4.3.2.1 Multiple search function	25
4.3.2.2 Accurate search result	25
4.3.2.3 Retrieve latest Electronic Thesis and Dissertation	25
4.3.2.4 Multiple components	25
4.3.3 Software requirement	25
4.3.4 Hardware requirement	26
4.4 Summary	26
5. Chapter 5 : System Implementation	27
5.1 Introduction	27
5.2 Implementation hierarchy model	27
5.3 Implementation of System Module	32
5.3.1 Web Server	32
5.3.2 Database tools	32
5.3.3 Web Browser	32
5.4 Technology used	32
5.4.1 Java Servlet	33
5.4.2 Java	34
5.4.3 Hyper Text Markup Language (HTML)	34
5.5 System Environment	34
5.6 User Interface for the process of student submission Electronic Thesis and Dissertations	35
5.6.1 Log In	35
5.6.2 Student Submission Form	35
5.6.3 Upload form	37

5.7 Conclusion	37
6. Chapter 6 : Experimentation	38
6.1 Introduction	38
6.2 Problem Encountered	38
6.2.1 Development Tools	38
6.2.2 Combination of Tools	38
6.2.3 System Debugging	38
6.3 Experimentations	38
6.3.1 The feature size which give the optimal classification accuracy	39
6.3.2 The effective similarity threshold value for the similarity measurement	39
6.4 Result Analysis	39
6.5 System limitations	40
6.6 Summary	40
7. Chapter 7 : Conclusion and Future Works	41
7.1 Introduction	41
7.2 Achievement	41
7.3 Future work	41
7.3.1 Converter	41
7.3.2 Accurate Index Terms List	41
7.3.3 All Categories in Electronic Thesis and Dissertations System	41
7.4 Conclusion	42
Bibliographies	43
Appendices	
A. Gantt Chart	44
B. Index terms list	46

List of Tables

Table 2.1 Comparisons of ETD functions	9
Table 4.1 Categories in the automated text classifier	17
Table 6.1 Precision and recall result	39

List of Figures

Figure 1.1 Automated text classifier system	3
Figure 1.2 Process of automated text classifier of electronic thesis and Dissertations	4
Figure 1.3 Electronic Thesis and Dissertation diagram	5
Figure 2.1 Electronic Thesis Database of University of Waterloo	10
Figure 2.2 Electronic Thesis Submission Upload form of University of Waterloo	11
Figure 2.3 Broad outline of an information retrieval system	12
Figure 3.1 Seven phases for automated text classifier of electronic thesis and dissertations	17
Figure 4.1 shows major tasks that are involved in the prototype system	20
Figure 4.2 Index file format	22
Figure 5.1 User selection	27
Figure 5.2 Document vector	28
Figure 5.3 Similarity metric	29
Figure 5.4 Category of Electronic Thesis and Dissertations system	29
Figure 5.5 Electronic Thesis and Dissertations titles	30
Figure 5.6 Retrieve similar electronic thesis and dissertations	31
Figure 5.7 Automated classify new electronic thesis and dissertations	31
Figure 5.8 Java Servlet coding	33
Figure 5.9 HTML code	34
Figure 5.10 User Log In page	35
Figure 5.11 Student submission form (Students information)	36
Figure 5.12 Student submission form (Account Information)	36
Figure 5.13 Student Upload form	37

Abstract

With the explosively growth of electronic devices, the demanding needs to provide scholarly materials available electronically has become part of the learning culture. Many universities are currently in the process of digitizing their theses and dissertations in an effort to preserve it and to make it more accessible to anyone.

This thesis describes a framework for the development of an information retrieval system to retrieve the similar electronic thesis and dissertations and to automate the process of classifying electronic thesis and dissertations. The previous electronic thesis and dissertations will be used as an input to a classifier system. A similarity coefficient, Jaccard coefficient is used to generate similarity metric for a collection of past years electronic thesis and dissertations. Similarity metric is used to measure the document similarity between electronic thesis and dissertations report. These similarities metric will then used to classify new unseen electronic thesis and dissertations into relevant categories.

We have developed a prototype system that defines a framework for the retrieval of relevant electronic thesis and dissertations and automatic categorization of electronic thesis and dissertations. The prototype system comprises of the document vector, document similarity and similarity metric. We used precision and recall to measure the quality of the results returned by the system.

The framework for the automated text classifier provide user precise relevant electronic thesis and dissertations according to user query and able to automatic classify an unseen electronic thesis and dissertations.

Abstrak

Memandangkan pertumbuhan yang mendadak dalam penggunaan peralatan elektronik, ia telah meningkatkan permintaan untuk mendapatkan bahan akademik secara elektronik. Pada masa sekarang, kebanyakan universiti sedang dalam proses untuk menjadikan tesis dan bahan akademik diperolehi secara elektronik. Ini dapat menolong orang ramai untuk memperolehi bahan akademik secara lebih mudah.

Tesis ini akan membandingkan kesamaan metrik yang mengukur kesamaan dokumen antara laporan tesis elektronik dengan menggunakan kesamaan coefficient, "Jaccard coefficient". Sistem prototaip ini terdiri daripada vektor dokumen, kesamaan dokumen dan kesamaan. Keupayaan dan kebolehan sistem kesamaan dokumen ini diuji dengan membandingkan kesemua laporan tesis elektronik dengan sistem tesis elektronik.

Kami telah membina satu system prototaip yang menerangkan satu rangka kerja untuk mendapatkan dan mengautomasi maklumat thesis elektronik dan disertasi dalam pelbagai kategori. Prototaip system tersebut menggabungkan vector dokumen, persamaan dokumen dan persamaan metric. Kami menggunakan ketetapan dan panggilan untuk menilai kualiti keputusan yang dikembalikan oleh system.

Rangka kerja untuk mengklasifikasikan teks secara automasi membolehkan pengguna mendapatkan maklumat thesis elektronik dan disertasi dengan tepat mengikut kehendak maklumat pencarian pengguna. Selain itu, ia juga mampu mengklasifikasikan thesis atau disertasi elektronik yang belum pernah dilihat secara automatik.

Chapter 1 Introduction

1.1 Introduction

Approximately 60,000 university students graduate every year. Theses and dissertations provide tangible evidence of the scholarly development of students, their ability to discover and effectively communicate research findings. ETDs stand for Electronic Thesis and Dissertations system. It defined as theses and dissertations submitted, archived, or accessed primarily in electronic formats.

Students were required to submit their theses and dissertations in electronic form rather than in paper. Electronic Thesis and Dissertations system is expressed in a form simultaneously suitable for machine archives and worldwide retrieval. The ETD is similar to its paper predecessor. It has figures, tables, footnotes, and references. It has a title page with the authors' name, the official name of the university, the degree granted, and the names of the group members. It documents the author's years of academic commitment (Anonymous, 2002a). It describes the problem statements, the objectives, the benefits of the thesis, how the research relates to previous work as recorded in the literature review, the research methods used, the results, and the interpretation and discussion of the results, and a summary with conclusions.

The Electronic Thesis and Dissertations system is different in such a way that it provides a technologically advanced medium for expressing one's ideas. Student prepares their thesis and dissertations using nearly any word processor or document preparation system, incorporating relevant multimedia objects.

Currently users may search for the electronic thesis and dissertations by category, author name, year of acceptance and title but there is no similar document retrieval function for electronic thesis and dissertations. Moreover, that is not an efficient way to search for the relevant electronic thesis and dissertations. Information retrieval technique can retrieve the similar electronic thesis and dissertations to user by using similarity metric. The process of how to generate the similarity metric will be described in this thesis.

Thesis and dissertations submitted are manually classified by human into correct categories. This process of classification is repeated for subsequent years. Current research in information retrieval could provide technologies to automatic classify the electronic thesis and dissertations. The work described in this thesis focuses on the automated classification of electronic thesis and dissertations.

This thesis will describe the process of a text classifier to automate the process of classifying electronic thesis and dissertations. The archived electronic thesis and dissertations will be used as an input to a classifier system. A similarity metric will be used to automatically compare the similarity between a collection of past years electronic thesis and dissertations classified by human. These derived classifiers could then used to classify new unseen incoming electronic thesis and dissertations into relevant categories.

At the same time, Thesis and dissertations submitted are usually manually classified by human into each relevant category. This process of classification is repeated for subsequent years. Current research in information retrieval could provide technologies to automatic classify the electronic thesis and dissertations. The work described in this thesis focuses on the automated classification of electronic thesis and dissertations.

1.2 Problem Statement

Conventionally, theses and dissertations are submitted, distributed and stored in paper form. However, numerous problems such as handling and storage of paper in the library and the graduate office are time-consuming and costly. The graduate office pays binding and filming costs, and must manually maintain an online index. The student/author pays high photocopy fees. All those problems are associated with the existing procedures for submission, distribution and storage of theses.

Currently, user retrieved the electronic thesis and dissertations by browsing the category in electronic thesis and dissertations system. Each category hosts the theses based on faculty program genre such as networking, computational science, and software engineer. This does not provide insight for user if an interested this is relevant to other theses. Indeed, the relevant electronic thesis and dissertations user required may not come from the same category. In this case, the search for relevant documents must be exhaustive and inaccurate because user may obtain a number of non-relevant documents.

Beside, the assignment of theses category is based on the natural language content. Conventionally, Electronic Thesis and Dissertations system used human intervention to classify the electronic thesis and dissertations. However, as the number of text increases, it becomes difficult for human to consistently classify them.

1.2.1 Problem solution

Electronic submission, distribution, and storage of theses and dissertations will address these problems. A number of well-known universities such as West Virginia University, University of Virginia, and University of Singapore have now begun to accept electronic submission of theses and dissertations, and then post these documents on library web servers.

Electronic theses and dissertations are positive development in graduate education on several counts: graduate student research achieves broader exposure, open up new opportunities for creative scholarships, students receive some experience in the technological skills required by many scholars today. At the same time, institution and student reduce the costs involved in publishing and storing dissertations and theses.

In order to overcome the retrieval of relevant electronic thesis and dissertations from different category to user, similarity coefficient is used for similarity measures to measure the similarity between the vectors. We will develop an automated text classifier system to perform electronic thesis and dissertations classification. Figure 1.1 illustrates the architecture of the prototype system for automated classification of electronic thesis and dissertations. This automated text classifier comprises of document vector, similarity measurement and the similarity metric.

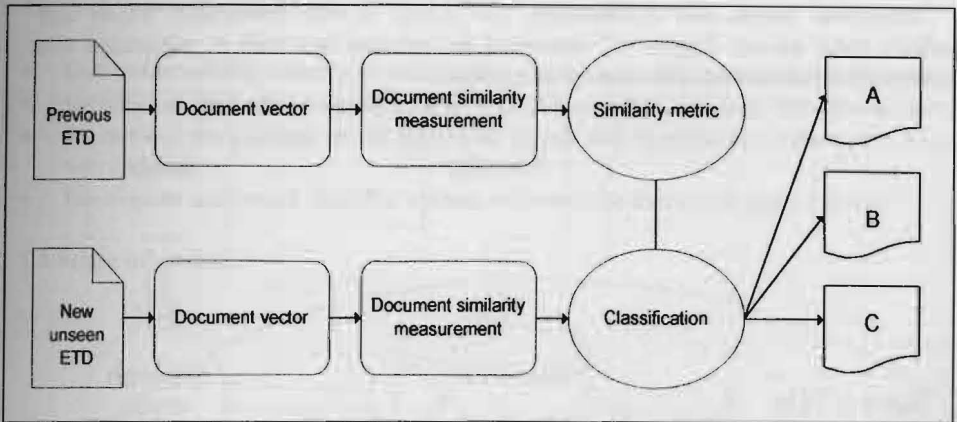


Figure 1.1 Automated text classifier system

For document vector, consider a collection of documents in which each document is characterized by one or more index terms. Thus, the documents are the objects in the collection each of which is represented by a number of properties (here index terms). The similarity between two objects is normally computed as a function of the number of properties that are assigned to both objects; in addition, the number of properties that is jointly absent from both the objects may also be taken into account. Substantially similar methods can be used for determining collection structure (by comparing pairs of text vectors with each other and identifying text pairs found to be sufficiently similar), and for retrieving information (by comparing the query vectors with the vectors representing the stored items and retrieving items that are found to be similar to the queries).

Traditionally, human was asked to judge every possible pair of text units, and mark them as similar or not similar. Similarity measures value increases when the number of common properties in two vectors increases.

Similarity metric is defined as co-occurring terms are projected onto the same dimensions. Information retrieval is based on the similarity match between the both documents in the electronic thesis and dissertations measured based on a similarity metric, instead of the exact match techniques used in traditional classification systems. Since it is difficult to design a similarity metric, which exactly conforms to human perception, it is likely that some items determined to be relevant or similar to the documents by the system are actually not relevant to the documents judged by the user. For this system, similarity metric takes value in $[0,1]$.

The top five highest similarity measures will then be used to predict the category of incoming electronic thesis and dissertations. Thus, past year electronic thesis and dissertations will be used to generate similarity metric and forthcoming electronic thesis and dissertations will be used for classification usage.

1.3 Automated text classifier for Electronic Thesis and Dissertations system

Electronic thesis and dissertations are stored in the Electronic Thesis and Dissertations local server. Figure 1.2 described the process involved in automated text classifier system of electronic thesis and dissertations.

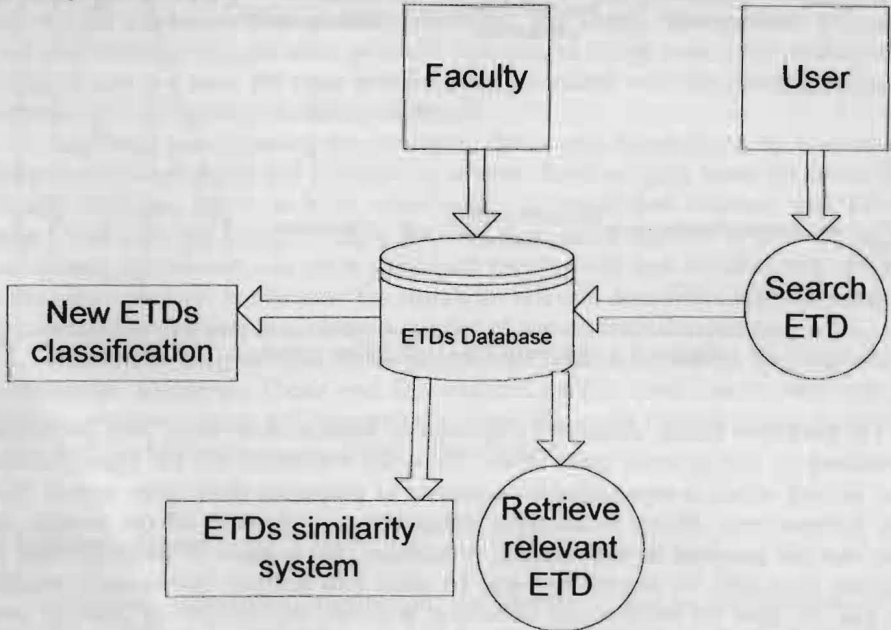


Figure 1.2 Process of automated text classifier of electronic thesis and dissertations.

We are implementing an automated text classifier of electronic thesis and dissertations which the prototype system used similarity metric to retrieve relevant electronic thesis and dissertations to the user, performing new electronic thesis and dissertations classification, users able to search electronic thesis and dissertations and finally we evaluate the electronic thesis and dissertations retrieval result.

The automated text classifier system will be put in the faculty. The faculty staff has the authority to use the electronic thesis and dissertations similarity system and responsible for the new electronic thesis and dissertations classification.

1.4 Objectives

The central aim of the Electronic Thesis and Dissertations system is to establish a distributed database of digital theses produced by graduate students at the participating institutions with the theses accessible on the web. The objective of this research is to explore the utility of a fully automated system for electronic thesis and dissertations. This study will enable us to define an efficient text classifier system for automated electronic thesis and dissertations classification system.

There are 4 primary objectives:

- Determine optimal number of precise index terms used to represent the documents.
- Develop an appropriate similarity metric to measure the similarity between vectors.
- To retrieve the relevant set of electronic theses and dissertations effectively based on user request.
- Develop an automated classifier system of electronic thesis and dissertations.

1.5 Scope of project

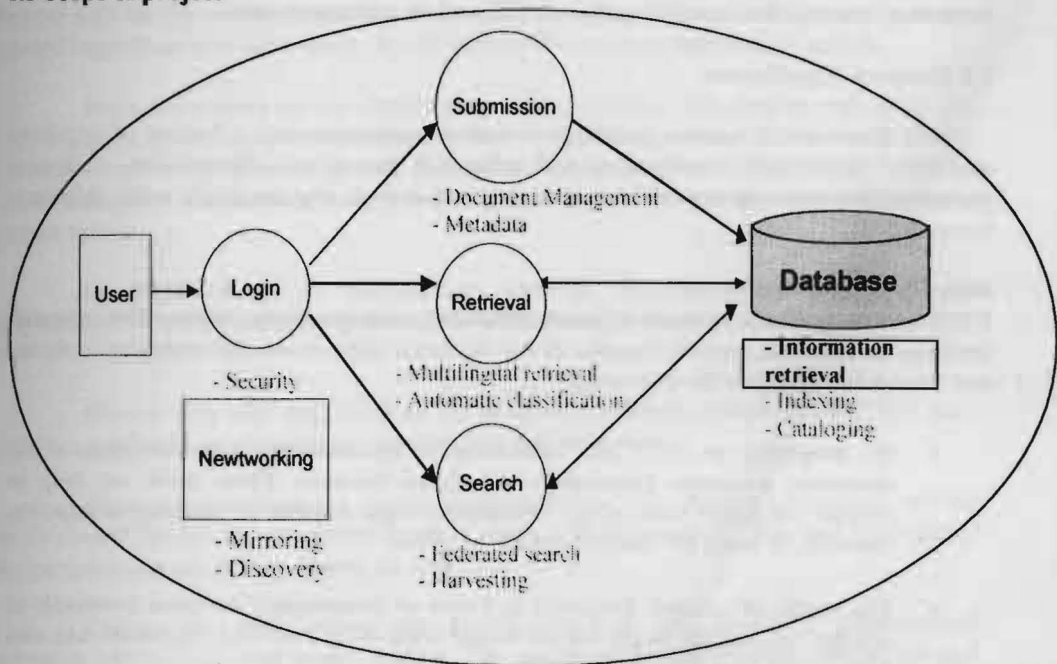


Figure 1.3 Electronic Thesis and Dissertation diagram

Diagram above shows possible research areas that have been identified for the Electronic Thesis and Dissertations system. The project focuses on information retrieval technique to help organize the database. This is to further enhance the accessibility of large dataset and retrieve the relevant data effectively through sorting the data to identify patterns and establish relationships. Research and identification algorithms, approaches in pattern matching or to determine the key relationship in the data stored in database are essential in developing the document similarity system. The final output would be a running prototype that can customize the database into more organized and structured database.

I will develop an automated text classifier system to perform document similarity retrieval function and automatic classify the new coming electronic thesis and dissertations. The recent developments in search engine are changing the world fast; the need for accurate result in information retrieval has become more important. Figure 1.2 illustrates the

architecture of the prototype system for automated classification of electronic thesis and dissertation. Information retrieval has been applied in the process of classification to perform document text processing. An information retrieval system informs the user of the existence and whereabouts of documents relating to his request. An information retrieval system is able to analyze the contents of the sources of information as well as the users' request, and then match these to retrieve those items that are relevant. The system involves porter stemming, index terms matching, document vector, perform similarity coefficient for similarity measures between vectors. Finally, the similarity measures will generate the similarity metric and store in a table and documents comparison will be performed for all datasets. Electronic thesis and dissertation will work as documents set.

1.6 Research Significance

There are, of course, problems as well as opportunities in allowing or requiring electronic submission, distribution, and achieving theses and dissertations, including problems of access and distribution, achieving and storage, and copyright and publication issues.

Below is the benefit of ETD:

ETDs enable graduate students to more effectively and creatively present their research findings. Some of the specific benefits of the electronic submission and archiving of theses and dissertations include the following:

- By preparing an ETD and submitting it electronically, one can learn about electronic document preparation and digital libraries. These skills can help to prepare for future roles in the Information Age, whether in teaching, conducting research, or using the research results of others.
- The results of research presented in theses or dissertations are more accessible to scholars all over the world via the World Wide Web. Potential employers may also more easily view these documents.
- The message of a thesis or dissertation may be better conveyed in an electronic as opposed to a paper document. Creative possibilities are expanded by allowing color diagrams, color images, hypertext links, audio, video, animation, spreadsheets, databases, simulations, etc., to be integrated into a document.
- By submitting theses electronically, the university is able to fulfill more economically its responsibilities of recording and archiving theses and dissertations. This is a key responsibility of the University library, and is easier and less costly to fulfill when the workflow involves electronic documents. Electronic Thesis and Dissertations system reduce the need for library storage space and advance digital library technology, which improves library services.

1.7 Project Plan

The project is started at 8th of July, 2002. The project documentation is divided into two parts. First part is project introduction, literature review, methodology, and system design. The second part will be focus on system implementation, experimentation, result

and analysis. The project will be completed at 31th of January, 2003. Please refer to Appendix A for more information.

1.8 Outline of Project Report

This report will be divided into seven chapters. Every chapter will define different scope of works involved in developing the system.

In chapter one, a brief introduction and problem statement will be discussed. This chapter will let the readers know the purpose of the study, scope of the project and the research significance of the project. It will explain the problem statement in details.

When the readers go into further observation, meaning that chapter two, he or she will obtain the reviewing, penetrate the pro and con of existing systems. Comparison among the technology, the programming used, the system features and the system interface will be drawn out. After the existing is reviewed, a new system will be proposed to improve the current system.

In chapter three, the methodology used to develop the prototype will be emphasized. Through the description of the methodology, the readers will understand the flow of the system.

Chapter four will emphasize on the proposed prototype system design. The user, software and hardware requirement specification will be defined.

Chapter five defines the system implementation. The designed prototype will be implemented through the programming part, which is to be compiled. The user interface of the prototype system will be display as well.

Chapter six will involve the experimentations. The proposed experimental approach will be described in this chapter. The results of the experiments carried out and the performance of the automated classify the 10 categories for Electronic Thesis and Dissertations are highlighted.

Finally, in chapter seven, it concludes the work and recommends the future works. It describes the improvement that might be developed to enhance the prototype system.

Chapter 2 Literature Review

2.1 Introduction

In this chapter, we will go through the introduction of information retrieval system. In order to have the correct view and comparison, some Electronic Thesis and Dissertations system and automated classification system have been browsed through to have a complete investigation of the existing systems.

User may request to retrieve the similar electronic thesis and dissertations but the relevant electronic thesis and dissertations are not necessary from the same category. Currently, Electronic Thesis and Dissertation system does not have the similarity documents retrieval technique, which is able to retrieve the similar electronic thesis and dissertations from different category, which can response to the user request. To address the problem, I propose an automated text classifier system for electronic thesis and dissertation. It also presents the methods used to evaluate the effectiveness of automated text classifier.

2.2 History

In this section, we will review the history of Electronic Thesis and Dissertations system, information retrieval and text categorization system to understand how the ETD system, information retrieval and text categorization system were developed. We will briefly describe the history of ETD system, information retrieval and text categorization system

2.2.1 Electronic Thesis and Dissertation system

The Electronic Thesis and Dissertation Project (ETD) was launched in 1987 at an Ann Arbor meeting arranged by UMI and attended by representatives of Virginia Tech, the University of Michigan, SoftQuad, and ArborText. Virginia Tech funded the development of a Document Type Definition (DTD) for dissertations and theses; SoftQuad's Yuri Rubinski wrote the initial DTD.

Virginia Tech's Electronic Thesis and Dissertation Project is part of the larger National Digital Library of Theses and Dissertations (Anonymous, 2002e), funded by a grant from the U.S. Dept. of Education. Approximately 15 other universities in supporting this effort have joined VT. The Southeastern Universities Research Association (SURA) has also provided funding for the ETD Project (Janet Erickson, 1997).

2.3 Reviewing of existing system

Reviewing of existing system is essential to understand the current system as well as to propose a better system to overcome the existing system problems and shortages. I had reviewed the existing Electronic Thesis and Dissertation system, comparison of the Electronic Thesis and Dissertation system function and the Text Categorization system.

2.3.1 Reviewing of Electronic and Dissertation System

Many countries have begun implementing Electronic Thesis and Dissertation project since 1987. Among the Electronic Thesis and Dissertation systems reviewed are:

2.3.1.1 Virginia Tech's Electronic Thesis and Dissertation System

This first ETD system, Virginia Tech's Electronic Thesis and Dissertation system is part of the larger National Digital Library of Theses and Dissertations (Anonymous, 2002d).

2.3.1.2 Networked Digital Library of Theses and Dissertations (NDLTD)

Several universities such as West Virginia University, University of Virginia, West Virginia University are members of the Networked Digital Library of Theses and Dissertations (NDLTD) in collaboration with other academic institutions. The NDLTD, originated at Virginia Tech, helps its member institutions share information in implementing ETD policies and ETD results (Anonymous, 2002e) The NDLTD presently has a total of 128 members, consisting of 113 member universities (including 3 consortia) and 15 institutions.

Among the universities website have been reviewed are:

- University of Waterloo (Christine Jewell, 2002).
- University of Virginia (Anonymous, 2002).
- West Virginia University (John Hagen, 2002).

Comparison has been drawn out among these websites. These websites are reviewed depending to the functions they provided.

2.3.2 Comparison of the functions

Comparison of ETD functions between three universities which are University of Waterloo, University of Virginia and West Virginia University is shown in table below.

Table 2.1 Comparison of ETD functions

Comparison	University of Waterloo	University of Virginia	West Virginia University
Student information form	/	/	/
Electronic Thesis upload form	/	/	/
Multimedia components	/	/	/
Searching	/	/	/
Password verification	/	/	/
Post script form	/		
Automated ETD classification			
Similar ETD retrieval			
Browse ETD with alphabetical			/

From the comparison table that is shown above, we realized that the three university electronic thesis and dissertations systems provided the basic function to user such as student information form, electronic thesis upload form, searching, multimedia components available and etc. The disadvantage of the three university electronic thesis and dissertations systems is that they cannot perform information retrieval technique and automated electronic thesis and dissertations classification. The searching for similar electronic thesis and dissertations must be exhausted and time consuming. Human intervention is needed because electronic thesis and dissertations system is unable to provide automated classification of electronic thesis and dissertations.

University of Virginia provides basic search and advance search for user to search for the Electronic Thesis and Dissertations.

UW Electronic Thesis Database

This site provides access to a selection of electronic theses and dissertations submitted by graduates of the University of Waterloo. These theses and dissertation have been accepted by the University as a partial requirement of a degree program at the Master's or PhD level.

One hundred and ten theses have now been added!

Search E-thesis Database by keyword, author, title, academic department, or year of acceptance

Enter search term or phrase:

(To search by full name, enter the name in the order "surname first name")

	Keyword ▾
Submit Query	Clear

Figure 2.1 Electronic Thesis Database of University of Waterloo.



University of Waterloo
Electronic Thesis Project

Etheses Submission Upload Form

Input the path and file name of your thesis or use the browse button to select the file on your hard drive. Please use your last name as your filename.

If you have supplementary files use the supplementary file boxes for submission.

Etheses File:	<input type="text"/>	<input type="button" value="Browse..."/>
Supplementary File 1:	<input type="text"/>	<input type="button" value="Browse..."/>
Supplementary File 2:	<input type="text"/>	<input type="button" value="Browse..."/>
Supplementary File 3:	<input type="text"/>	<input type="button" value="Browse..."/>
Supplementary File 4:	<input type="text"/>	<input type="button" value="Browse..."/>
Supplementary File 5:	<input type="text"/>	<input type="button" value="Browse..."/>
<input type="button" value="Upload!"/>		

Figure 2.2 Electronic Thesis Submission Upload form of University of Waterloo.

2.4 System Architecture

Below we briefly described the system architecture for information retrieval system.

2.4.1 Information Retrieval System

The term information retrieval was coined in 1952 and gained popularity in the research community from 1961 onwards. At that time information retrieval's organizing function was seen as a major advance in libraries that were no longer just storehouses of books, but also places where the information was catalogued and indexed (Chowdhury, 1999). Information retrieval techniques have been broadly used in text categorization task because both are content-based document management tasks (Cunningham, Litten and Witten, 1997; Korfhage, 1997).

The concept of information retrieval presupposes that there are some documents or records containing information that have been organized for easy retrieval. The documents or records we are concerned with contain bibliographic information that is quite different from other kinds of information or data. An information retrieval system has three major components that are the document subsystem, the user subsystem, and the searching/retrieval system. For electronic thesis and dissertations system, I am focusing on the searching/retrieval system.

2.4.1.1 Broad outline of an Information Retrieval system

An information retrieval system comprises six major sub-systems.

1. The document subsystem
2. The indexing sub system
3. The vocabulary sub system
4. The searching sub system
5. The user-system interface
6. The matching subsystem

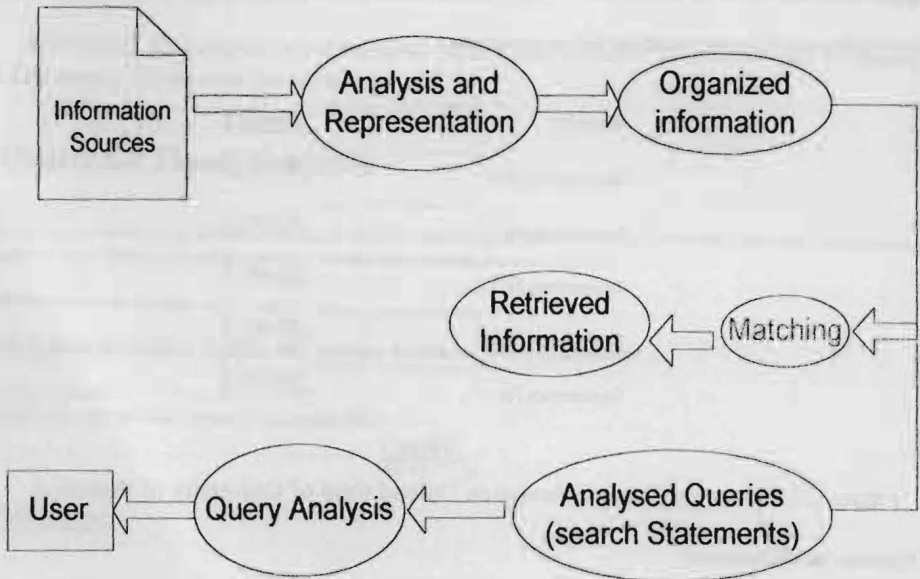


Figure 2.3 Broad outline of an information retrieval system

The entire task mentioned in Figure 2.3 can be brought under two major groups which are subject /content analysis, and search and retrieval. Subject or content analysis includes the task related to the analysis, organization and storage of information. The process of search and retrieval includes the tasks of analyzing users' queries, creation of a search formula, the actual searching, and retrieval of information.

2.4.1.2 Retrieval Models

It should be remembered that in most computerized information retrieval systems, the judgment as to whether a document is relevant or not to a given query is based on the topicality or lexical similarity between the query terms and the document space. Various mathematical models have been proposed to represent information retrieval systems and procedures. These include the Boolean search model, the Vector processing model and best-match searching. (Korfhage,1997)

2.4.1.2.1 Boolean search model

The Boolean search model, which compares Boolean query statements with the term set that is used to represent document contents; the probabilistic retrieval model, which is based on the computation of relevance probabilities for the documents of a collection.

Although Boolean searching has been used by almost all the information retrieval systems for quite some time, it has certain limitations. The first relates to the formulation of search statements. Users are not able to formulate an exact search statement by the combination of AND, OR, and NOT operators, especially when several query terms are involved. The second limitation relates to the number of retrieved items. Users cannot predict a priori exactly how many items are to be retrieved to satisfy a given query.

The last limitation of Boolean searching is that it identifies an items as relevant by finding out whether a given query term is present or not in a given record in the database. Thus, all retrieved items are considered to be of equal importance; however a given concept may be discussed in different documents with different emphasis or weight.

2.4.1.2.2 Vector Processing model

The vector processing model assumes that an available term set, called term vectors, is used for both the stored records and information requests. Collectively the terms assigned to a given text are used to represent text content.

Several coefficients for similarity measures can be used and there are a total of five coefficients. There are the Dice coefficient, the Jaccard coefficient, the Cosine coefficient, the Overlap coefficient and the Asymmetric coefficient (Salton and McGill 1983). All similarity measures exhibit one common property, namely that their values increase when the number of common properties (or the weight of the common properties) in two vectors increases. The Jaccard expression and the Cosine expression measures have similar characteristics, ranging from a minimum of 0 to a maximum of 1 for non-negative vector elements. Both these measures have been widely used for the evaluation of retrieval functions.

2.4.1.2.3 Best match searching model

Best match searching is designed to produce ranked output. It therefore requires a method to measure the relative importance of the retrieval items, which again requires some method of weighting the search terms.

A best match search matches a set of query words against the set of words corresponding to each item in the database, calculates a measure of similarity between the query and the item, and then sorts the retrieved items in order of decreasing similarity with the query.

2.4.1.3 Evaluation of an Information Retrieval System

In this information retrieval environment we may want to access how the level of performance of a given system can be improved. In 1996, Cleverdon had identified six criteria for the evaluation of an information retrieval system (Cleverdon, 1978). There are: