

UNIVERSITI TEKNOLOGI MALAYSIA

BORANG PENGESAHAN STATUS TESIS*

JUDUL: AN ANALYSIS OF HIERARCHICAL CLUSTERING AND NEURAL NETWORK CLUSTERING FOR SUGGESTION SUPERVISORS AND EXAMINERS

SESI PENGAJIAN : SEMESTER I 2005/2006

Saya NURUL NISA BINTI MOHD NASIR
(HURUF BESAR)

mengaku membenarkan tesis (PSM/Sarjana/Doktor—Falsafah)* ini disimpan di Perpustakaan Universiti Teknologi Malaysia dengan syarat-syarat kegunaan seperti berikut :

1. Tesis adalah hakmilik Universiti Teknologi Malaysia.
2. Perpustakaan Universiti Teknologi Malaysia dibenarkan membuat salinan untuk tujuan pengajian sahaja.
3. Perpustakaan dibenarkan membuat salinan tesis ini sebagai bahan pertukaran antara institusi pengajian tinggi.
4. ** Sila tandakan (✓)

SULIT

(Mengandungi maklumat yang berdarjah keselamatan atau kepentingan Malaysia seperti yang termaktub di dalam AKTA RAHSIA RASMI 1972)

TERHAD

(Mengandungi maklumat TERHAD yang telah ditentukan oleh organisasi/badan di mana penyelidikan dijalankan)

TIDAK TERHAD

Disahkan oleh



(TANDATANGAN PENULIS)



(TANDATANGAN PENYELIA)

Alamat Tetap:

JA 8139, BATU 22,
KAMPUNG BUKIT SENGGEH,
77500 SELANDAR, JASIN,
MELAKA.

PM DR NAOMIE BINTI SALIM
Nama Penyelia

Tarikh : 30 NOVEMBER 2005


Tarikh : 30 NOVEMBER 2005

CATATAN: * Potong yang tidak berkenaan.

** Jika tesis ini SULIT atau TERHAD, sila lampirkan surat daripada pihak berkuasa/organisasi berkenaan dengan menyatakan sekali sebab dan tempoh tesis ini perlu dikelaskan sebagai SULIT atau TERHAD.

♦ Tesis dimaksudkan sebagai tesis bagi Ijazah Doktor Falsafah dan Sarjana secara penyelidikan, atau disertasi bagi pengajian secara kerja kursus dan penyelidikan, atau Laporan Projek Sarjana Muda (PSM).

“We hereby declare that we have read this thesis and in our opinion this thesis is sufficient in terms of scope and quality for the award of the degree of Master of Science (Computer Science)”

Signature : 

Name of Supervisor : .PM DR NAOMIE BINTI SALIM

Date : 30 NOVEMBER 2005

**AN ANALYSIS OF HIERARCHICAL CLUSTERING AND NEURAL
NETWORK CLUSTERING FOR SUGGESTION SUPERVISORS AND
EXAMINERS**

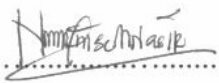
NURUL NISA BINTI MOHD NASIR

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

NOVEMBER 2005

I declare that this thesis entitled “*An Analysis Of Hierarchical Clustering And Neural Network Clustering For Suggestion Supervisors And Examiners*” is the result of my own research except as cited in the references. The thesis has not been accepted for any degree and is not currently submitted in candidature of any other degree.

Tandatangan :.....
Nama Penulis : NURUL NISA BINTI MOHD NASIR
Tarikh : 30 NOVEMBER 2005

Especially for Ayah & Mama...

Thanks for your love, guidance and blessings...

For my brothers & sisters... Along, Alang, Ateh & Adik,...

You are the best in the world...

For my Man...whom always there to support me...

Thanks for everything...

ACKNOWLEDGEMENT

Profound acknowledgement and special thanks to supervisor, Associate Professor Dr. Naomie Salim for encouragement, critics and friendship.

Thanks to my family, especially my beloved dad and mom (I know you are hearing me), for understanding and encouraged to continue my studies.

Not forgetting also to my college and friends for their support and assistance in helping me to achieve this major milestone of my life. Lastly, my thanks to Herman, for the continuous support throughout the years.

ABSTRACT

Document clustering has been investigated for use in a number of different areas of information retrieval. This study applies hierarchical based document clustering and neural network based document clustering to suggest supervisors and examiners for thesis. The results of both techniques were compared to the expert survey. The collection of 206 theses was used and employed the pre-processed using stopword removal and stemming. Inter document similarity were measured using Euclidean distance before clustering techniques were applied. The results show that Ward's algorithm is better for suggestion supervisor and examiner compared to Kohonen network.

ABSTRAK

Dewasa ini, kaedah pengelompokan dokumen banyak diaplikasikan dalam bidang Capaian Maklumat. Kajian ini akan mengadaptasikan pengelompokan dokumen berasaskan Rangkaian Neural dan juga Pengelompokan Timbunan Berhirarki. Hasil pengelompokan ini dianalisis bagi mencari kaedah terbaik dalam pemilihan penyelia dan penilai dan dibanding dengan pemilihan yang dilakukan oleh pakar. Dokumen-dokumen yang dikelompokkan menjalani pra-pemprosesan termasuklah penghapusan perkataan yang tidak membawa makna dan mempunyai kekerapan yang tinggi atau *stopword*, pembuangan imbuhan atau *stem*, dan seterusnya pengelompokkan kata nama supaya tiada pengulangan perkataan yang sama. Seterusnya, keserupaan dokumen-dokumen selepas pra-pemprosesan akan digambarkan menggunakan jarak Euclidean. Hasil yang diperolehi menunjukkan algoritma Ward's adalah lebih baik dalam pemilihan penyelia dan penilai berbanding algoritma Kohonen.

TABLE OF CONTENT

CHAPTER	TITLE	PAGE
	ABSTRACT	v
	ABSTRAK	vi
	TABLE OF CONTENT	vii
	LIST OF TABLES	xi
	LIST OF FIGURES	xii
	LIST OF ABBREVIATION	xiii
	LIST OF SYMBOL	xiv
	LIST OF APPENDICES	xv
1	INTRODUCTION	1
	1.1 Introduction	1
	1.2 Problem Background	2
	1.3 Problem Statement	5
	1.4 Objectives	5
	1.5 Project Scope	5
	1.6 Significance of Project	6
	1.7 Organization of Report	6

2	LITERATURE REVIEW	8
2.1	Introduction	8
2.2	Background of Assigning Supervisors and Examiners in FSKSM	8
2.3	Information Retrieval	9
2.4	Text Preprocessing	10
	2.4.1 Stopword Removal	10
	2.4.2 Stemming	11
	2.4.3 Noun Groups	12
	2.4.4 Index Term Selection	12
	2.4.5 Indexing	13
2.5	Document Representation	15
2.6	Document Clustering	16
	2.6.1 Hierarchical Clustering	19
	2.6.2 Kohonen Clustering	24
2.7	Clustering Performance Measure	28
2.8	Discussion	28
2.9	Summary	30
3	EXPERIMENTAL DETAIL	31
3.1	Introduction	31
3.2	Thesis Collection and Digitization	32
3.3	Stopword Removal	33
3.4	Stemming	33
3.5	Document Vector Representation	34

3.6	Data Sampling	35
3.7	Ward's Clustering	36
3.7.1	Euclidean Distance	36
3.7.2	Combining RNN and Ward's Clustering	36
3.7.3	Mojena's Stopping Rule	37
3.8	Kohonen Clustering	39
3.8.1	PCA Implementation	39
3.8.3	Kohonen Network Algorithm	40
3.9	Evaluation of Ward's Clustering and Kohonen Clustering Compared to Expert Survey	42
3.10	Summary	43
4	RESULTS AND ANALYSIS	44
4.1	Introduction	44
4.2	Preprocessing Result	44
4.3	Evaluation of Ward's Clustering and Kohonen Network	45
4.3.1	Ward's Result	45
4.3.2	Kohonen Result	46
4.4	Comparative Study and Discussion	48
4.5	Summary	50
5	CONCLUSION	52
5.1	Summary	52
5.2	Contribution	53
5.3	Further Work	53

REFERENCES	55
APPENDIX A	63
APPENDIX B	66
APPENDIX C	72
APPENDIX D	77
APPENDIX E	80
APPENDIX F	86
APPENDIX G	88
APPENDIX H	90
APPENDIX I	103

LIST OF TABLES

TABLE NO	TITLE	PAGE
2.1	Time and space complexity of several well known algorithms	22
2.2	Web/Document clustering in previous research	28
3.1	Splitting sample	35
3.1	Kohonen network design	41
4.1	Ward's cluster	91
4.2	Ward's prediction - Sample 50:50	92
4.3	Ward's prediction - Sample 60:40	95
4.4	Ward's prediction - Sample 75:25	98
4.5	Ward's prediction - Sample 80:20	100
4.6	Ward's prediction - Sample 95:5	102
4.7	Ward's prediction	46
4.8	Kohonen prediction - Sample 50:50	104
4.9	Kohonen prediction - Sample 60:40	107
4.10	Kohonen prediction - Sample 75:25	110
4.11	Kohonen prediction - Sample 80:20	112
4.12	Kohonen prediction - Sample 95:5	114
4.13	Kohonen prediction	47
4.14	Comparative study	48

LIST OF FIGURES

FIGURE NO	TITLE	PAGE
2.1	Text pre-processing	10
2.2	Hierarchical clustering dendrogram	19
2.3	Basic algorithm of RNN	21
2.4	Complete linkage	23
2.5	Kohonen network architecture	25
2.6	Kohonen network weight	25
2.7	Competitive learning networks	26
2.8	Weight adjustment	27
3.1	Framework of study	32
3.2	Ward's algorithm	39
4.1	Accuracy of Ward's algorithm	46
4.2	Accuracy of Kohonen algorithm	48
4.3	Ward's Performance vs Kohonen Performance	49

LIST OF ABBREVIATION

ANN	- Artificial Neural Network
BMU	- Best Matching Unit
HAC	- Hierarchical Agglomerative Clustering
IR	- Information Retrieval
IRS	- Information Retrieval System
NN	- Neural Network
PCA	- Principal Component Analysis
RNN	- Reciprocal Nearest Neighbour
SOM	- Self Organizing Map
STC	- Suffix Tree Clustering

LIST OF SYMBOL

S_{D_i, D_j}	-	Similarity between document i and document j
$weight_{ik}$	-	k -th weight in document i
$weight_{jk}$	-	k -th weight in document j
ESS	-	Error sum squares
η	-	Learning rate

LIST OF APPENDICES

APPENDIX	TITLE	PAGE
A	Gantt Chart	63
B	Thesis Collection	66
C	Stopword List	72
D	Porter Stemming Rule	77
E	Preprocessing Result	80
F	Supervisor Code	86
G	Expert Code	88
H	Ward's Performance	90
I	Kohonen Performance	103

CHAPTER 1

INTRODUCTION

1.0 Introduction

IR is a discipline involved with the organization, structuring, analysis, storage, searching and dissemination of information. A compact definition of the basic function of an *information retrieval system (IRS)* has been given by Lancaster, (1968):

“An information retrieval system does not inform (i.e. change the knowledge of) the user on the subject of his enquiry. It merely informs on the existence (or non-existence) and whereabouts of documents relating to his request.”

Much of the research and development in IR is aimed at improving the effectiveness and efficiency of retrieval. Document clustering was introduced to IR on the grounds of its potential to improve the efficiency and effectiveness of the IR process. Jardine and Van Rijsbergen (1971) provided some experimental evidence to suggest that the retrieval efficiency and effectiveness of an IR application can benefit from the use of document clustering. The efficiency and effectiveness of an IR application was expected to increase through the use of clustering, since the file organization and any strategy to search it, take into account the relationships that hold between documents in a collection (Croft, 1978). Relevant documents that might have otherwise been ranked low in a best-match search will be (through inter-

document associations) grouped together with other relevant documents, thus improving the efficiency and effectiveness of an IR system.

1.2 Problem Background

Document clustering has been applied to IR for over thirty years. The aim of research in the field is to postulate the potential of clustering to increase the efficiency and effectiveness of the IR process (Jardine & Van Rijsbergen, 1971; Croft, 1978). The literature published in the field covers a number of diverse areas, such as the visualization of clustered document spaces (Allen *et al.*, 2001; Leuski, 2001), the application of document clustering to browsing large document collections (Cutting *et al.*, 1992; Hearst & Pedersen, 1996), etc.

The main motivation for this work has been to investigate methods for the improvement of the efficiency and effectiveness of document clustering. One type of clustering employed in this study is hierarchical clustering; perhaps the most commonly used type of clustering in IR (Willett, 1988). This is a choice based on the more sound theoretical basis of hierarchical clustering. Jardine and Sibson (1971), Salton and Wong (1978) and Van Rijsbergen (1979) have identified three strengths of hierarchical methods. Firstly, such methods are theoretically attractive since they do not depend on the order in which documents are processed. Secondly, they are well formed, in the sense that a single classification will be derived from a given set of documents. And finally, hierarchic methods are stable, since small changes in the original document vectors will result in small changes in the resulting hierarchies.

The application of hierarchical methods to IR (e.g. group average, complete link and Ward's methods) was extensively investigated during the 1980s. The majority of the research work was carried out at Cornell University by Voorhees (1985a) and at Sheffield University by Griffiths *et al.* (1984, 1986) and also El-Hamdouchi and Willett (1989).

More recently, information science researchers have turned to other newer artificial intelligence based inductive learning techniques including neural networks. This newer techniques which are grounded on diverse paradigms have provided great opportunities for researchers to enhance the information processing and retrieval capabilities of current information storage and retrieval systems.

NN is another clustering technique applied in this study. Neural network models have many attracting properties and some of them could be applied to an IR system. Recently, there is a research tendency to apply NN in cluster document. Initially, Kohonen is an unsupervised NN which is mathematically characterized by transforming high-dimensional data into two dimensional representations, enabling automatic clustering of the input, while preserving higher order topology.

In neural network models, information is represented as a network of weighted, interconnected nodes. In contrast to traditional information processing methods, neural network models are "self-processing" in that no external program operates on the network: the network literally processes itself, with "intelligent behavior" emerging from the local interactions that occur concurrently between the numerous network components (Reggia & Sutton, 1988). It is expected that the research on the application of neural network models into IR will grow rapidly in the future along with the development of its technological basis both in terms of hardware and software (Qin He, 1999).

Neural networks computing, in particular, seem to fit well with conventional retrieval models such as the vector space model and the probabilistic model. Doszkocs et al. (1990) provided an excellent overview of the use of connectionist models in IR. A major portion of research in IR may be viewed within the framework of connectionist models.

Essentially, thesis focuses solely on the retrieval effectiveness and efficiency of document clustering for suggestion of supervisors and examiners for thesis since there is not much research in this domain.

Each Computer Science student enrolled in the master program should produce a thesis before finishing his/her studies. This thesis contains a complete report of a research.

Each thesis should have at least one supervisor and examiners to fulfil the requirement. This supervisor and examiners are selected either from FSKSM lecturers or any other person in who is expert in the thesis's subject.

A supervisor is responsible to guide student in doing research, and producing a valuable research whereas examiners evaluate the yield of research and to see whether students really understand his/her research. The evaluation from the supervisor and examiners shows the quality of a student's research.

Currently, the determination of supervisor and examiner is done manually by the coordinators. However, sometimes the coordinators are new and did not know much about the experience of lecturers in supervising and examining students in various areas. The selection process based on incomplete knowledge such as this sometimes may affect the quality of thesis produced by students. The major problem related to thesis performance is the student didn't get an effective guidance from his/her supervisor because the supervisor is not the expert in the thesis's subject.

The weaknesses of such a manual system may affect the quality of research in the long term.

Therefore, in this study, two clustering techniques are used, Kohonen clustering and Hierarchical clustering to give a better solution. Clustering result will be analyzed in order to find out the best techniques for the solution. Furthermore the implementation of mathematical algorithm makes the system more concrete without bias situation.

1.3 Problem Statement

- Can document clustering be used for determining supervisors and examiners of thesis effectively?
- Can Kohonen based document clustering perform better result than Ward's clustering (one type of hierarchical clustering) for determining supervisors and examiners of thesis?

1.4 Objectives

The objective of this study is as follows:

1. To represent index terms in document vector.
2. To apply two techniques of clustering, Kohonen clustering and Ward's clustering to improve the efficiency and effectiveness of suggestion for supervisors and examiners
3. To analyze Kohonen network based document clustering and Ward's based document clustering for suggestion supervisors and examiners
4. To compare clustering techniques to use in the domain of suggestion of supervisors and examiners in FSKSM, UTM

1.5 Project Scope

Two clustering techniques will be applied in this study that is Neural Network clustering and Hierarchical clustering. The result of these two clustering will be analysed to find out the best techniques in domain study. This study will be done in scope as stated below:

1. Title and abstract of 206 theses will be stored on the machine and will be used in information retrieval process. The theses are on master thesis from FSKSM, UTM only.
2. Porter stemming will be used to reduce a word to its *stem* or root form in the title and also the abstract of thesis
3. Indexing process will create a unique identifier of the documents by counting the frequency of each index terms before the *tfidf* weighting is calculated
4. Ward's clustering and Kohonen clustering will be applied to the indexed documents.

1.6 Significance of the Project

Results of the study will show whether NN based document clustering or Hierarchical based document clustering is effective for determining supervisors and examiners. It will also give insight on whether NN based is better than Hierarchical based document clustering in terms of suggestion of supervisors and examiners.

1.7 Organization of the Report

This report consists of five chapters. The first chapter presents introduction to the project and the background of problem on why is the study is being conducted. It also gives the objectives and scope of the study. Chapter 2 reviews on IR, pre-processing to achieve IR purpose, and document clustering also clustering techniques that will be used in this study. Chapter 3 discusses on the framework of this project in detailed including pre-processing phase further clustering algorithm that will be

applied in this study. Chapter 4 contains a cluster analysis based on Ward's and Kohonen performance in determining supervisors and examiners and Chapter 5 is the conclusion and suggestions for future work.

CHAPTER II

LITERATURE REVIEW

2.1 Introduction

This chapter will review previous research related to this study. Previous research is very useful in order to have a good research. This review will touch a little bit of information retrieval, document clustering also pre-processing of document to create an indexed file for document clustering; which are stemming and stopword removal.

2.2 Background of Assigning Supervisors and Examiners in FSKSM

Each Computer Science student in master program should produce with one thesis before finishing his/her studies. This thesis should contain a complete report of research. For this purpose, each student will be supported and guided by at least one supervisor. Each report will be evaluated by some examiners. The selection of examiner is determined during the postgraduate committee meeting.

In this research, we are concerned with the selection of supervisors and examiners. This is an important step in order to get a better report and outcome in the studies. The selection process has to be carried out appropriately because it will affect the thesis performance.

In the past, students will choose their own supervisors. The aim is to fulfil the requirement of project I without much consideration of the expertise and knowledge of the supervisor. Consequently, some supervisors can't guide students in producing a good research results since they have no experience in such kind of work in a particular area.

There are also examiners had to evaluate student research and improper selection can results in choices that are not accurately based on their expertise and knowledge. The main issue is the ability of the examiners in evaluating research methodology and result.

The selection and approval process is normally done by coordinators and postgraduate committee. However, they sometimes do not have enough information on the experience of the lecturers to do the selection effectively. The selected supervisor and examiner should have knowledge and capability in term of proposed studies.

This situation shows some weaknesses of human intervention in constructing decision especially if they are inexperienced. Therefore this study tries to find out the best solution to the problem using mathematical algorithm, with the hope to improve the efficiency and effectiveness of selection supervisors and examiners.

2.3 Information Retrieval

Information retrieval is essentially a matter of deciding which documents in a collection should be retrieved to satisfy a user's need for information. Thus, an information retrieval aims at collecting and organizing information in one or more

subject areas in order to provide it to the user (Salton and McGill, 1983). IR techniques have been developed over many years to support searching for documents (Van Rijsbergen, 1979; Baeza-Yates & Ribeiro-Neto, 1999). There is much research in using IR techniques (Dunlop, 2000).

2.4 Text Pre-processing

Yet the goal of text pre-processing is to optimise the performance of data analysis such as clustering. The first step of most machine learning algorithms is to reduce dimensionality by discarding irrelevant data. Data analysis is very dependent on the pre-processing and the data representation model. This is the most important step before implementing document representation further similarity measures. Figure 2.1 briefly illustrated the text pre-processing in order to produce index term towards representing document. The followed paragraph will explain briefly the pre-processing in Fig. 2.1.

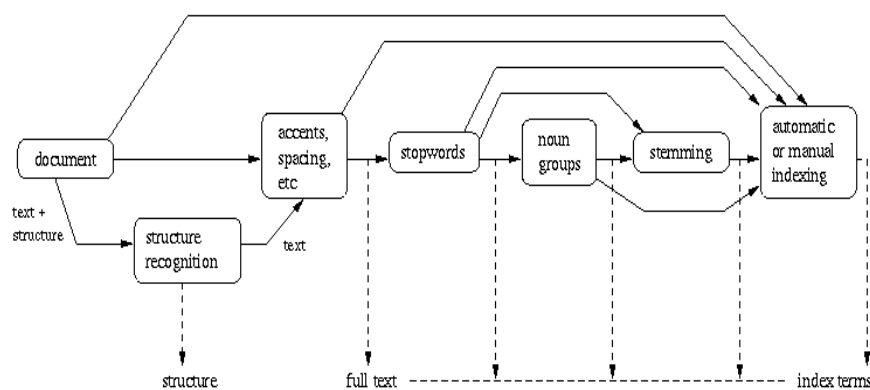


Figure 2.1 Text pre-processing

2.4.1 Stopword Removal

A text retrieval system often associates a stop list with a document set, which is a set of words that are deemed “irrelevant”, e.g., *a, the, of, for, with*, etc., even though they may appear frequently. Stoplists may vary when document set varies,

e.g., “computer”. This kind of term is removed by compiling stopword lists so they do not interfere with the data analysis.

2.4.2 Stemming

Word stemming is a process of text normalizations which consists of reducing individual words, or words within phrase structures, to some abstract (canonical) representation. For example, the words: “presentation”, “presented”, “presenting” could all be reduced to a common representation “present”. Stemming has been the most widely applied morphological technique for information retrieval.

Stemming also reduces the total number of distinct index entries. Further, stemming causes query expansion by bringing word variants and derivation (Pirkola, 2001). Some early research results with English collections questioned the effectiveness of stemming (Harman, 1991).

There are two kinds of stemming errors which are understemming errors, in which words which refer to the same concept are not reduced to the same stem, and overstemming errors, in which words are converted to the same stem even though they refer to distinct concepts. In designing a stemming algorithm there is a trade-off between these two kinds of error. A light stemmer plays safe in order to avoid overstemming errors, but consequently leaves many understemming errors. A heavy stemmer boldly removes all sorts of endings, some of which are decidedly unsafe, and therefore commits many overstemming errors.

A number of techniques have been proposed in the past (Frakes and Baeza-Yates, 1992). In all these methods, the individual rules are accompanied by conditions designed to improve the accuracy of the stemming and to prevent words from being shortened too far (e.g., to prevent “ring” and “red” being converted to “r”).

A well-known technique for stemming text is Porter's algorithm, which is based on a set of rules extracted from the English language (Porter, 1980). Specifically it has five steps applying rules within each step. Within each step, if a suffix rule matched to a word, then the conditions attached to that rule are tested on what would be the resulting stem, if that suffix was removed, in the way defined by the rule. Since Porter stemmer is a very widely used and available stemmer, and is used in many applications, this stemmer is employed to the collection of 206 thesis in this study.

Another stemmer is Paice/Husk stemmer, a simple iterative stemmer whereas it removes the endings from a word in an indefinite number of steps (Paice, 1994). Lovins stemmer is a single pass, context-sensitive, longest-match stemmer (Lovins, 1968). This stemmer, though innovative for its time, has the problematic task of trying to please two masters (IR and Linguistics) and cannot excel at either. The Dawson stemmer is another stemmer which it is a complex linguistically targeted stemmer that is strongly based upon the Lovins stemmer, extending the suffix rule list to approximately 1200 suffixes (Dawson, 1974). It keeps the longest match and single pass nature of Lovins, and replaces the recoding rules, which were found to be unreliable. The other side of stemmer is Krovetz stemmer which effectively and accurately removes inflectional suffixes in three steps (Krovetz et. al., 1993).

2.4.3 Noun Group

The index terms tend to be very large so terms that are similar and close to each other are mapped to one term via word stemming.

2.4.4 Index Term Selection

Index term selection goal is to reduce the number of words in the vector description. There are numerous methods for keyword selection such as by extracting

keywords based on their entropy. In the approach discussed by Klose et. al., (2000), for each word k in the vocabulary the entropy as defined by Lochbaum and Streeter (1989) was computed:

$$W_k = 1 + \frac{1}{\log_2 n} \sum_{j=1}^n P_{jk} \log_2 P_{jk} \quad \text{with} \quad P_{jk} = \frac{tf_{jk}}{\sum_{l=1}^n tf_{lk}}$$

where tf_{jk} is the frequency of word k in document j , and n is the number of documents in the collection.

Borgelt and Nurnberger (2001) applied a greedy strategy for index term selection where all the terms obtained after text pre-processing will then categorized in the selected term (index term) with highest relative entropy. This process can be terminated until the desired number of index term has been selected.

Due to objective of this study, index term selection will not performed and consequently, all the term produces after text pre-processing will then indexed to employ the two clustering techniques as described in Chapter 1.

2.4.5 Indexing

Indexing is the use of language to describe the documents and user's information needs. Index terms are derived from the text or the user input. An *indexing term* is defined as a set of unique words that characterize some feature found in the document set. Documents are encoded or modeled using the indexing terms that appear in them. User queries are processed by mapping the query to the indexing terms previously extracted from the document set and then matching the query to individual documents. Indexing is done either by human experts or an automatic indexing program. Manual document indexing is labour-intensive and time consuming work and has a drawback of lack of uniformity and indexer user mismatch. In contrast, automatic indexing has the advantage of bias free uniformity and efficiency.

Automatic indexing usually proceeds by the following steps: (i) get words in each document, (ii) exclude stop words from them, (iii) do stemming to produce index terms, (iv) compute the precoordination information (e.g. term frequency) and pointers from the term to documents to build an inverted file.

The goal of indexing is to build a data structure that will allow quick searching of the text. There are many classes of indices based on different retrieval approaches (e.g. inverted file, signature file and so on). Almost all type of indices are based on some kind of tree or hashing.

Increasingly, IR databases are designed to provide more than one indexing approach in hopes of maximizing the effective retrieval of useful messages, texts and documents. This study will performed automatic indexing for document representation.

Documents were indexed using:

- i) *tf/idf* weighting, which weights terms proportional to how often they occur in the current document but inversely to how often they occur in the collection as a whole (Sparck Jones, 1972)
- ii) a simple stopword list based on the collection itself, the 30 most common words in the collection were not indexed
- iii) Porter stemming algorithm, an algorithmic stemmer that conflates variants of a words into the same base form, e.g. *walking*, *walks* etc all conflate to *walk* (Porter, 1980)
- iv) Cosine matching function, an IR standard that takes into account term weights and document lengths (Salton and McGill, 1983).

The IR engine was designed to index the corpus, in this study only title and abstract will be indexed before the *tfidf* weighting is calculated.

2.5 Document Representation

For IR purpose, we have to map the text files to numerical feature vectors. There exist numerous models in document representation. The top three models are Boolean, Vector space and probabilistic models.

The probabilistic models, involve estimating probabilities. The goal is to estimate the probability of relevance of a given document to a user with respect to a given query. Probabilistic assumptions about the distribution of elements in the representations within relevant and irrelevant document are required (Maron & Kuhns, 1960).

The author has employed the vector space model for document representation (Salton, 1989). In the vector space model, term weights can be interpreted as probability estimates (Turtle, 1992) and a great deal of experimental work has been done to evaluate alternative forms (Salton & Buckley, 1988). In general, these are referred to as *tf/idf* weights, since they include a component based on the frequency of a word (or feature) in the text of an object (the term frequency component or *tf*) and a component based on frequency of the word in the “universe” of objects (the inverse document frequency of *idf*). The *idf* weight increases as the frequency of the word decreases (hence the name). For example;

For a given term w_j and document d_i

$$tf_{ij} = \frac{n_{ij}}{|d_i|} \quad idf_j = \frac{\log n_j}{n} \quad x_{ij} = tf_{ij} \bullet idf_j$$

n_{ij} is the number of occurrences of w_j in document d_i

$|d_i|$ is the number of words in document d_i

n is the number of documents

n_j is the number of documents that contain w_j

*x_{ij} is the *tfidf* for term w_j in document d_i*

The cosine similarity is the most commonly used method in vector space model to compute the similarity between two documents d_i and d_j , which is defined

to be $\cos(d_i, d_j) = \frac{d_i^t d_j}{\|d_i\| \|d_j\|}$. The cosine formula can be simplified to

$\cos(d_i, d_j) = d_i^t d_j$ when the document vectors are of unit length. This measure becomes one if the documents are identical and zero if there is nothing in common between them (e.g., the vectors are orthogonal to each other).

In particular, Boolean model purpose is to find documents satisfying the query in Boolean form. However the model has the limitation according to Cater and Craft (1987) and Wong, et.al, (1988).

2.6 Document Clustering

Clustering techniques have long been used in IR to improve the performance of search engines both in term of timing and quality or results [e.g. Jardine and Van Rijsbergen, 1971; Van Rijsbergen and Croft, 1975; Griffiths, et. al. 1986). This work follows from the observation, known as the cluster hypothesis, that relevant documents are more like one another than they are to non-relevant documents (Van Rijsbergen & Spark Jones, 1973).

Clustering means that documents in collection are processed and grouped into dynamically generated clusters. Dunlop (2000) investigated the use of clustering techniques to improve the performance of people matching based on web.

Document clustering goal is to automatically group related documents based on their content. This technique requires no training sets or predetermined taxonomies and generates taxonomy at runtime. To be able cluster text document collections with hierarchical and NN clustering, therefore, the author first to applied pre-processing methods e.g., stopword removal, stemming, encoded each document

using vector space model and finally selected a subset of terms as features for the clustering process.

Choosing the variables and similarity measurements is the first step in a cluster analysis. This is a very important step, since the decision on these issues will affect the final results directly (Willett, 1988). At this step, the raw data matrix will be converted in to a matrix of inter-individual similarity (dissimilarity or distance) measures. Some clustering methods have their specific measure (e.g. Euclidean distance for Ward's method) but more commonly the choice of measure is at the discretion of the researcher.

Most common measures that are widely used are Dice coefficient, Jaccard coefficient and Cosine coefficient.

Dice coefficient:

$$S_{D_i, D_j} = \frac{2 \sum_{k=1}^L \text{weight}_{ik} \text{weight}_{jk}}{\sum_{k=1}^L \text{weight}_{ik}^2 + \sum_{k=1}^L \text{weight}_{jk}^2}$$

Jaccard coefficient:

$$S_{D_i, D_j} = \frac{\sum_{k=1}^L \text{weight}_{ij} \text{weight}_{jk}}{\sum_{k=1}^L \text{weight}_{ik}^2 + \sum_{k=1}^L \text{weight}_{jk}^2 - \sum_{k=1}^L (\text{weight}_{ik} \text{weight}_{jk})}$$

Cosine coefficient:

$$S_{D_i, D_j} = \frac{\sum_{k=1}^L (\text{weight}_{ik} \text{weight}_{jk})}{\sqrt{\sum_{k=1}^L \text{weight}_{ik}^2 \sum_{k=1}^L \text{weight}_{jk}^2}}$$

In particular, distance measures have enjoyed widespread popularity because intuitively they appear to be dissimilarity measures. Distance measures normally have no upper bounds and are scale-dependent. Euclidean distance is defined as

$$D(V_x, V_y) = \frac{\sqrt{\sum_{j=1}^k (V_{xj} - V_{yj})^2}}{n} \quad (1)$$

where n is the number of documents, k is the number of elements in vectors V_x and V_y . V_{xj} is the j th component of the V_x vector.

The Euclidean distance takes the difference between two representation directly. It should therefore only be used for expression data that are suitably normalized, for example by converting the measured representation levels to log-ratios. In the sum, we only include terms for which both x_i and y_i are present, and divide by n accordingly. Unlike the correlation-based distance measures, the Euclidean distance takes the magnitude of changes in the gene expression levels into account.

Clustering performed more effective than simple searching for basic IR techniques (Dunlop, 2000). The standard clustering algorithm can be categorized into hierarchical algorithms such as single linkage, complete linkage, average linkage or Ward's clustering and partitioning algorithm such as k-means.

More recently, information science researchers have turned to other newer artificial intelligence based inductive learning techniques including neural networks. This newer techniques which are grounded on diverse paradigms have provided great opportunities for researchers to enhance the information processing and retrieval capabilities of current information storage and retrieval systems.

Since there are numerous clustering algorithms, the author had focus on hierarchical clustering and NN clustering. Hierarchical clustering is widely used in document clustering research (e.g. Dunlop, 2000, Leuski, 2001) because of its effectiveness and the quality of cluster produced.

2.6.1 Hierarchical Clustering

A hierarchical clustering algorithm creates a hierarchy of clusters. This type of structure is particularly useful in IR as it allows a document collection to be viewed at different levels of graininess. It builds a tree where each node is a cluster of objects and the clusters corresponding to the node's immediate children form a complete partition of that cluster (Mirkin, 1996). On input the algorithm receives a set of objects and a matrix of inter-object distances. It starts by assigning each object to its own unique cluster that is the leaves of the future tree. The algorithm iterates through the cluster set by selecting the closest pair of clusters and merging the together forming a new cluster that replaces them in the cluster set. A node corresponding to this new cluster is created in the tree and the selected pair of clusters becomes its children. That procedure is executed until all objects are contained within a single cluster, which becomes the root of the tree. Fig. 2.2 shows the dendrogram for hierarchical clustering. This is a general algorithm that is instantiated.

The clustering techniques were used to produce a hierarchical clustering of the thesis, *H*, that hopefully, has similar users grouped together on the lower levels of the hierarchy.

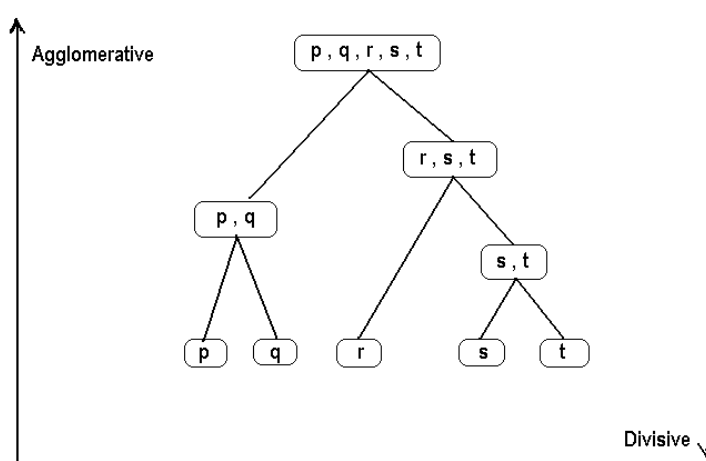


Fig. 2.2 Hierarchical clustering dendrogram

Initially, there are four well known techniques in hierarchical clustering which are single linkage, average linkage, group linkage and Ward's method.

Ward's Method

Ward's method on the other hand aims to merge clusters that result in the minimum loss of information which results from the grouping of objects into clusters and to quantify that loss in a form that is readily interpretable, such as the total sum of squared deviations of every object from the mean of the cluster to which it belongs. At each step in the analysis, union of every possible pair of clusters is considered and the two clusters whose fusion results in the minimum increase in the sum of the distances from each object to the centroid of its clusters.

When a method satisfies the reducibility property (Murtagh, 1983), a more computationally efficient reciprocal nearest neighbour (RNN) algorithm that can produce exact results as those produced by the original clustering method can be used. The main advantage of this algorithm is that all reciprocal nearest neighbours can be simultaneously merged, without affecting the final dendrogram. The reducibility property requires that if the following distance constraints hold for clusters i, j and k for some distance ρ : $d(i,j) < \rho$, $d(i,k) > \rho$ and $d(j,k) > \rho$, if the agglomerated cluster is $(i+j)$, then $d(i+j,k) > \rho$. It also implies that when clusters i and j are merged into clusters $(i+j)$, we have to update the nearest neighbour only for those points which have i or j as nearest neighbours. The Ward's method and the group-average method (implemented using the Cosine coefficient) satisfy this property, thus they can be implemented using the RNN algorithm. This algorithm traces a path through the similarity space until a pair of points is reached that are both more similar to one another than they are to any other points. They are called reciprocal nearest neighbours (RNN) and are combined to form a single new point. The search for other RNNs continues until all points have been combined. The basic RNN algorithm is shown in Figure 2.3.

Basic algorithm of RNN can be described as below:

1. Mark all entities as unfused
2. Starting at an unfused I , trace a path of unfused nearest neighbours (NN) until a pair of RNNs is encountered, i.e., trace a path of the form $J := NN(I)$, $K := NN(J)$, $L := NN(K)$... until a pair is reached for which $Q := NN(P)$ and $P := NN(Q)$.
3. Add the RNNs P and Q to the list of RNNs along with the distance between them. Mark Q as fused and replace the centroid of P with the combined centroid of P and Q .
4. Continue the NN-chain from the point in the path prior to P , or choose another unfused starting point if P was a starting point.
5. Repeat Steps 2 to 4 until only one unfused point remains.

Fig. 2.3 Basic algorithm of RNN

Based on Table 2.1, we can see that combination ward's and RNN gives best result among other hierarchical clustering techniques. The combination of Ward's and RNN is proposed by El-Hamdouchi and Willet (1986). RNN which is proposed by Murtagh (1983) can improve the time complexity of Ward's technique.

As we can see in Table 2.1, hierarchical clustering (first five rows) shows its effectiveness in term of space complexity. Although its time complexity shows low performance, but hierarchical clustering still maintain the quality of clusters produced (Steinbach, et. al., 2000; Dunlop, 2000; Mock, 1998).

Hierarchical Clustering Technique	Time Complexity	Space Complexity
Single Linkage	$O(N^2)$ - (Sibson, 1973)	$O(N^2)$ Van Rijsbergen (1971), (Sibson, 1973)
Complete Linkage	$O(N^2)$ - (Defays, 1977)	$O(N^2)$ - (Defays, 1977)
Average Linkage	$O(N^2)$ - Voorhees (1985a, 1986)	$O(N)$ - Voorhees (1985a, 1986)
Ward's	$O(N^3)$ - (Ward, 1963)	$O(N^2)$ - (Ward, 1963)
Ward's +RNN	$O(N)$ - (El-Hamdouchi, Willet, 1986)	$O(N^2)$ - (El-Hamdouchi, Willet, 1986)
K-means	$O(N)$	$O(N^2)$
Leader	$O(N)$	$O(N^2)$
ISODATA	$O(N)$	$O(N^2)$

Complete Linkage

Also known as the '*furthest neighbour*' method since it measures the distance between two groups as the most distance pair of individual objects, one from each group. The parameters for complete linkage are: $aX = 0.5$, $aY = 0.5$, $b = 0$ and $g = 0.5$. Which gives:

$$D_{P,Q} = \frac{D_{X,Q}}{2} + \frac{D_{Y,Q}}{2} + \frac{|D_{X,Q} - D_{Y,Q}|}{2}$$

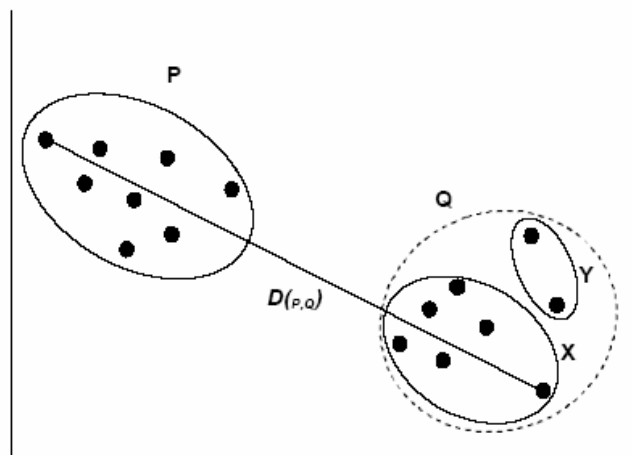


Fig. 2.4 –Complete linkage;. The new cluster Q is formed from combining the two groups X and Y .

Single Linkage

The single-link algorithm is more versatile than the complete-link algorithm, otherwise. For example, the single-link algorithm can extract the concentric clusters but the complete-link algorithm cannot. However, from a pragmatic viewpoint, it has been observed that the complete link algorithm produces more useful hierarchies in many applications than the single-link algorithm (Jain and Dubes 1988).

Group-Average

The group-average method measures distance by taking the average of the distances between all pairs of individual objects from the two groups. The parameters for group-average are:

$$\alpha_X = \frac{N_X}{N_P}, \quad \alpha_Y = \frac{N_Y}{N_P}, \quad \beta = 0 \text{ and } \gamma = 0,$$

which gives

$$D_{P,Q} = \frac{N_X D_{X,Q}}{N_P} + \frac{N_Y D_{Y,Q}}{N_P}$$

N_X and N_Y are the number of objects in the clusters X and Y respectively. Also, $N_P = N_X + N_Y$.

2.6.2 Kohonen Clustering

Neural networks computing, in particular, seems to fit well with conventional retrieval models such as the vector space model (Salton, 1989) and the probabilistic model (Maron & Kuhns, 1960). Kohonen network, a specific kind of ANN, is a tool that may be used for the purpose of automatic document categorization (Kohonen, 1997). This model is actually called a self-organising map, or SOM an unsupervised competitive ANN. The aim of a SOM is to produce a pattern classifier which is self-organising, using Kohonen learning to adjust the weights.

Lin et al. (1991) had adopted a Kohonen network for information retrieval. Kohonen's feature map, which produced a two dimensional grid representation for N -dimensional features, was applied to construct a self-organizing map (unsupervised learning), visual representation of the semantic relationships between input documents. In MacLeod and Robertson (1991), a neural algorithm was used for document clustering.

Typically, a Kohonen network consists of a 2-dimensional array of neurons with all of the inputs arriving at all of the neurons. Each neuron j has its own set of weights which can be thought of as together representing a "prototypical pattern" which is "remembered" by that neuron. When an input pattern arrives at the network, the neuron with the prototype pattern which is most similar to the input pattern will give the largest response, thus "recognizing" the input pattern. The key defining property of Kohonen is that the prototype patterns are stored in such a way that similar prototypes are found in neurons that are physically close to each other, and prototypes that are very different from each other are situated far apart.

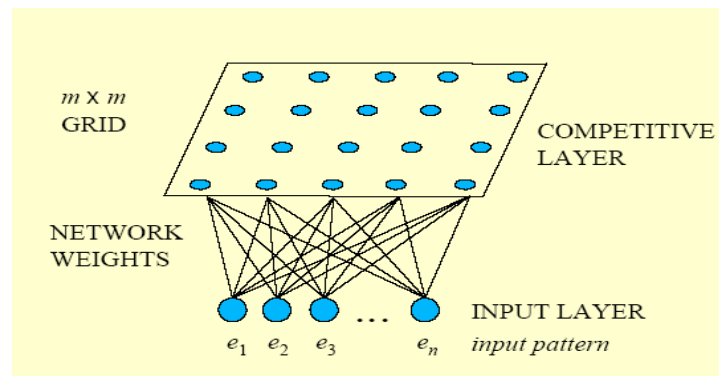


Fig. 2.5 Kohonen network architecture

With the demand for biological plausibility rising, the concept of self-organizing networks became a point of interest among researchers. Fig. 2.5 shows the architecture of Kohonen network. Self-organizing networks could be both supervised or unsupervised, and have four additional properties:

- Each weight is representative of a certain input (refer Fig. 2.6).
- Input patterns are shown to *all* neurons simultaneously.
- Competitive learning: the neuron with the largest response is chosen.
- A method of reinforcing the competitive learning.

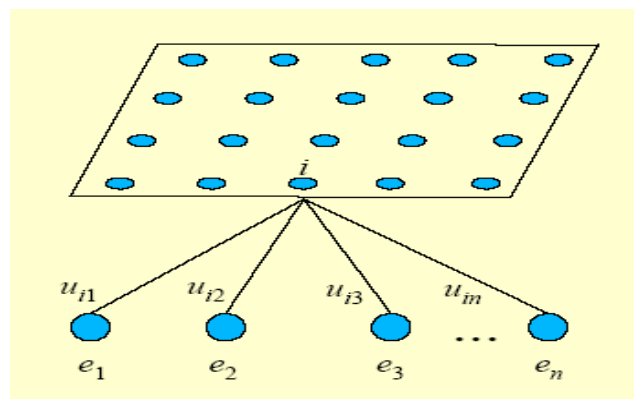


Fig. 2.6 Kohonen network weight

In competitive learning, neurons compete among themselves to be activated. In other words, only single output neuron is active at any time. The output neuron that wins the “competition” is called the *winner-takes-all* neuron. Fig. 2.7 shows the

competitive learning in Kohonen network. Competitive learning rule defines the change Δw_{ij} applied to synaptic weight w_{ij} as

$$\Delta w_{ij} = \begin{cases} \alpha(x_i - w_{ij}), & \text{if neuron } j \text{ wins the competition} \\ 0, & \text{if neuron } j \text{ loses the competition} \end{cases} \quad (1)$$

$$\alpha_t = \alpha_0 \left(1 - \frac{t}{T} \right)$$

Current iteration
Maximum iteration

where x_i is the input signal and α is the learning rate parameter.

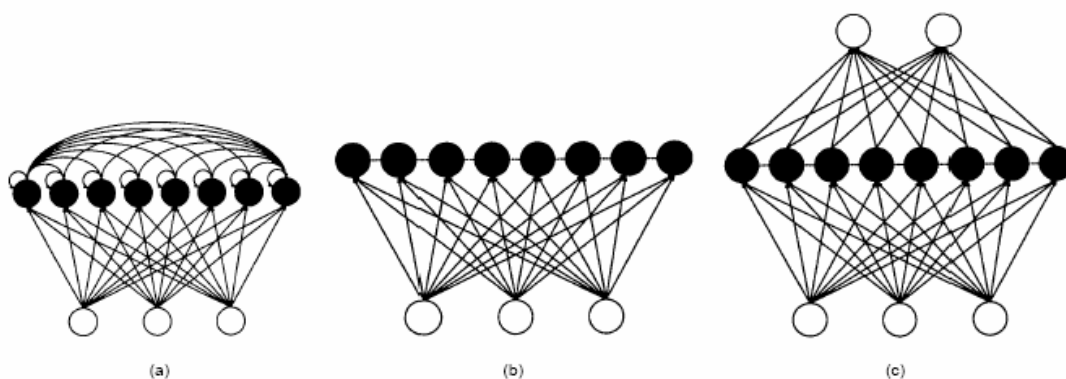


Fig. 2.7. Competitive learning networks (a) with explicit inhibitory connections among the competitive units, (b) with implicit inhibition, as in Kohonen's direct selection of the winner based on minimum distortion, and (c) with a supervised network above which employs inputs from the competitive layer. The competitive layer can be considered to perform either data compression or data representation by feature extraction from the input vectors.

At the initial process, the initialized weight is calculated randomly and in next iteration the weight is calculated as in equation (1) and the generated weight much better than the initialized weight (see Fig. 2.8).

The learning algorithm iterates until it converges (adjustments are arbitrarily

close to zero). Finally, each training document is mapped to a single node either through a simple matching of grid vectors to document vectors (Lin, et al.,1991) or by running an additional pass of the SOM algorithm to self organize a mapping (Honkela, et al., 1996).

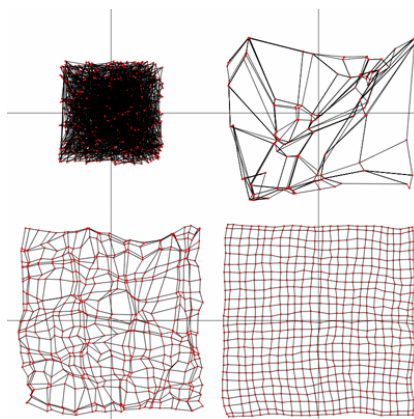


Figure 2.8 Weight adjustment: TL: Initial iteration, TR: 100 iterations, BL: 200 iterations, BR: 500 iterations.

The overall effect of the competitive learning rule resides in moving the synaptic weight vector W_j of the winning neuron j towards the input pattern X . The matching criterion is equivalent to the minimum Euclidean distance between vectors. The Euclidean distance between a pair of n -by-1 vectors X and is defined by

$$d = \|X - W_j\| = \left[\sum_{i=1}^n (x_i - w_{ij})^2 \right]^{1/2}$$

where x_i and w_{ij} are the i th elements of the vectors X and W_j , respectively.

The SOM transforms highly dimensional data into two dimensional grid, while keeping the data topology by mapping similar data items to the same cell on the grid (or neighbour cells). A typical SOM is made of a vector of nodes for the input, an array of nodes as the output map and a matrix of connections between each output unit and all the input units.

2.7 Clustering Performance Measure

Initially, prediction models such as ANN, aim to achieve an ability to correctly classify cases or problems unseen during training phase. Subsequently, the quality indicator is the accuracy during the testing phase. Generally, a classifier will be able to generalize if its architecture and learning parameters have been properly defined and enough training data are available.

In particular, the application of more than one data sampling may provide the basis for accurate and reliable predictions (Azuaje, 2003). Thus, the author had set-up five data samples as follows, 50:50, 60:40, 70:30, 80:20 and 95:5. The average of these five samples represents the accuracy for each of the prediction models.

Essentially, data sampling can be used to establish differences between data sampling techniques when applied to small and larger data sets, to study the response of these methods to the size and number of train-test sets also to discuss criteria for the selection of sampling techniques.

2.8 Discussion

Previous research in document clustering is discussed briefly in the following Table 2.2.

Table 2.2 Web/Document clustering in previous research

Author	Main Contribution
Kandel, et, al. (2003)	Proposed graph based k-means for web clustering
Leuski (2001)	Shows Ward's is an effective method for interactive information retrieval

Steinbach, Karpis and Kumar (2000)	Bisecting k-means shows the better performance than Kmeans
Na Tang and Rao Vemuri (2004)	AIS result is more compact cluster, good for large sized document sets that contain data redundancy
Korenius, et, al. (2004)	Ward's is the best method after stemming and lemmatization Finnish documents
Dunlop (2000)	Balanced Document clustering improve the performance of people matching based on web pages
Lin et al., (1991)	Adopted Kohonen for IR
Mock (1998)	Tree Cluster give good result in both domains, scales linearly with the input and generates a shallow tree hierarchy that may be easily browser
Botofago, (1993)	Proposed approaches rely solely on the semantic information embedded in link structures between documents (<i>link-based methods</i>)
Weiss et al., (1996)	Hybrid approach that combines link and content information in order to calculate interdocument similarities
Macskassy et al. (1998)	Conducted a small scale experiment to investigate the way that humans cluster web documents.
Zamir and Etzioni (1988)	Develop a clustering algorithm designed specifically for web documents (<i>Suffix Tree Clustering, STC</i>).

Based on Table 2.2, we can see that there are only a few researches on document clustering based on NN. Due to this observation, the author try to measure NN document clustering performance and compares it to the hierarchical clustering since it produced quality cluster among other algorithm.

Since Ward's clustering produce a quality cluster and its time and space efficiency than other hierarchical clustering techniques, the author decide to apply this technique to compared with Kohonen performance in suggestion supervisor and examiner. Essentially Kohonen clustering is the most popular technique in document clustering as described above. Accuracy percentages of both clustering algorithms will be counted to evaluate their performance in terms of suggestion of supervisors and examiners.

2.9 Summary

This chapter discussed the application of document clustering in IR. There is various applications were applied document clustering and we can see that Hierarchical clustering shows better performance in most application in terms of its quality cluster. However, more recently, most of the information science researchers have turned to other newer AI including NN. According to this study, document clustering was applied in determining supervisors and examiners in FSKSM wherein Ward's algorithm and Kohonen network has chosen to apply in this study.

CHAPTER III

EXPERIMENTAL DETAIL

3.1 Introduction

This chapter will explain the methodology used for this study. There are seven steps in the framework of this study. All the methodology is illustrated as in Fig. 3.1. This study is implemented using Intel Pentium 4 processor with 128 MB RAM and 40 GB of hard disk capacity. The software that was used in this study is Microsoft Visual C++, Visual Basic 6.0, Multivariate Statistical Package 3.13m (MVSP), also the Microsoft Office XP for report writing.

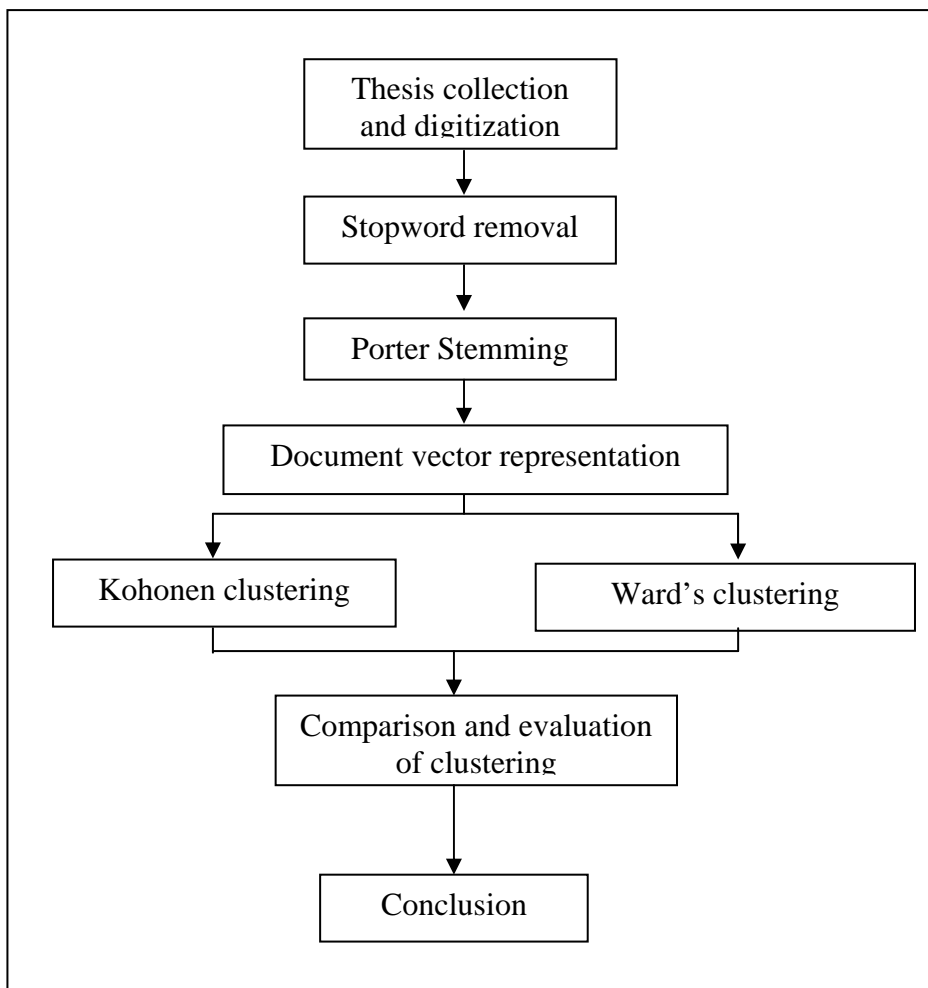


Figure 3.1 Framework of study

The next paragraphs will explain the details for each step. Scheduling and time management for this study is illustrated in the Gantt chart (refer to *Appendix A*).

3.2 Thesis Collection and Digitization

A set of thesis will be used in this study. This set contains 206 master theses from FSKSM, UTM. The theses are digitized for preprocessing of text before clustering. This entire document will be used as training data followed by clustering process based on samples that will be described in section 3.6. *Appendix B* shows the set of 206 theses that were used in this study.

3.3 Stopword removal

The first process is the stopword removal. As stated in the literature review, each document will contain an unneeded text or in other word the text is no significant in the next process. All this stopword can make storage became larger than supposed. Each text in each document (input text) will be compared with a stopword list. All this will execute using Visual Basic programming. The stopword list will be enclosed in *Appendix C*.

Initially, the list of stopwords will be defined. Then, every time a new document read, all the stopwords will be removed before proceeding to the next stage.

3.4 Stemming

Porter stemmer will be performed in order to remove the affix and successor suffix in thesis collection. Porter stemmer is chosen because this stemmer has been widely used in English stemming, in addition this study not focused on stemming effect.

There are five rules to be used in order to achieve goal of Porter stemmer. The rules are divided into steps. The rules in a step are examined in sequence, and only one rule from each step can be applied.

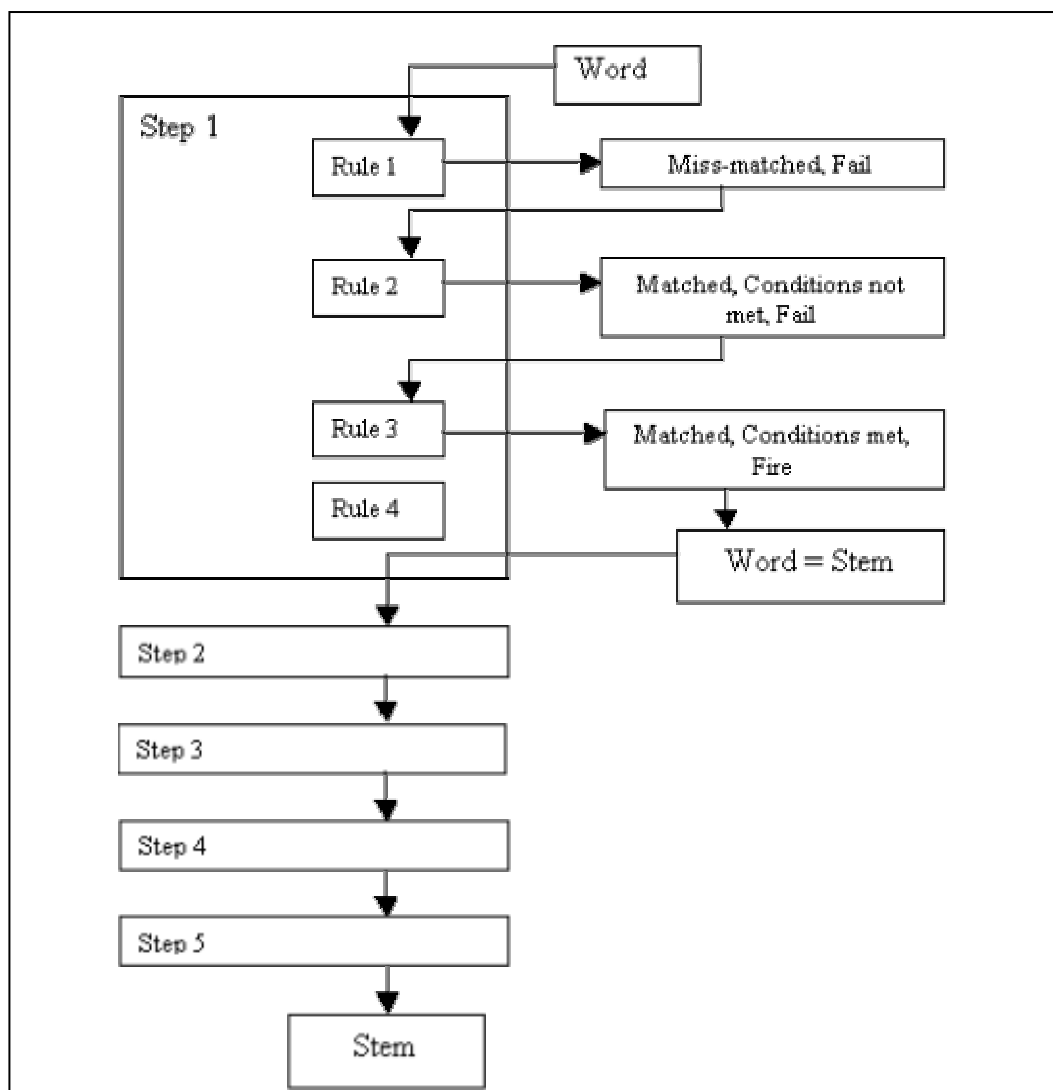


Fig. 3.2 Porter stemming

Refer *Appendix D* for detailed rule of Porter Stemming.

3.5 Document Vector Representation

Document vector is a process of representing the content of each document in a document collection so as to facilitate efficient and accurate retrieval of desired documents from the collection.

The algorithm used for this stage is as follows:

1. Determine index term after stopword removal and stemming
2. Store index term in array without duplicate term
3. Read document and match with index term array then store the frequency of term in document
4. Repeat step 3 for all document set

Document vector is one of the way to model the input for implementation in the clustering method. Each document and each query is represented by a vector or n -tuple. Each value represents a particular term produced after stopword removal and stemming process and the vector is in non binary form where the values represent the occurrences assigned to terms. For this study, the author used *tfidf* weight instead of occurrences for next clustering purpose. The *tfidf* weight is defined as in section 2.3.2.

3.6 Data sampling

As explained in section 2.5, data sampling is an alternative in order to measure the accuracy of prediction model. One of the data sampling techniques that are applied in this study is cross validation in which data will be divided randomly into the training and test sets. This process is repeated k times and the classification performance is the average of the individual test estimates.

The validation results were analysed for five different splitting methods as shown in Table 3.1.

Table 3.1 Splitting sample

Testing Set	Training Set
50% (1-103)	50% (104-206)
60% (1-82)	40% (83-206)
70% (20-70)	30% (1-19 and 71-206)
80% (1-42)	20% (43-206)
95% (148-157)	5% (1-147 and 158-206)

Sampling techniques can be implemented to assess the classification quality factors (such accuracy) of classifiers (such as ANNs) (Azuaje, 2003).

3.7 Ward's Clustering

Due to the time constraint, only one technique of hierarchical clustering is applied in this project and Ward's clustering has been chosen because it can generate quality cluster (Korenius, et. al., 2004; Leuski 2001). The document vector with *tfidf* weight which obtained after preprocessing data will be clustered using this technique.

3.7.1 Euclidean distance

Based on explanation in section 2.4, Euclidean distance is the distance measure that will be applied in implementing Ward's algorithm. This kind of measure is defined as in Equation (1), section 2.4.

3.7.2 Combining RNN and Ward's Clustering

The following algorithm shows the combination of Ward's clustering and RNN proposed by Murtagh (1983) which is more efficient:

- i. Initialize as much as possible clusters wherein each cluster contains exactly one document. At this stage, value for E is 0.
- ii. Combine RNN method in this step by getting the nearest neighbors (NN) for each cluster in consider the distance similarity

iii. Reduce the number of cluster by merging those two that minimize the increase of the total error sum of square, E using Equation (3) and those merged cluster will never isolate. At this stage, minimum variance is employed. The following equation is the calculation for error sum of square, ESS_k for cluster k

$$ESS_k = \sum_{i=1}^n x_{jk}^2 - \frac{1}{n} \left(\sum_{i=1}^n x_{jk} \right)^2 \quad (2)$$

where x_{jk} is an attribute value in document i which is clustered to cluster k sized n . Equation (3) represent as sum of the ESS for each cluster k defined as E and g refer to number of cluster.

$$E = g \sum_{k=1} ESS_k \quad (3)$$

At this step, two centroid clusters is chosen randomly to merged then updating error sum of square. The best combination of clusters will minimize total of the error sum of square.

iv. If there is more than one cluster remain, repeat step (iii)

Initially, time requirement for Ward's clustering algorithm $O(N^3)$. The combination of RNN proposed by Murtagh (1983), time requirement was reduced to $O(N^2)$ (El-Hamdouchi and Willett, 1989).

3.7.3 Mojena's stopping rule

In hierarchical clustering, the partition with the best number of groups for a dataset will need to be selected from the overall dendrogram. One partitioning

method is by examining the differences between fusion levels in the dendrogram [Mojena, 1977]. It can be implemented depends on the distribution of clustering criterion, α defined as

$$\alpha_p = \min_{i < j} [ESS_{ij}], i, j = 1, \dots, N - p \quad (4)$$

where α_{ij} is the value in which there is $N-p$ cluster. ESS_{ij} is depends on clustering technique used. For this study, Ward's clustering is used and the value of ESS_{ij} refer to the error sum of square obtained once merging cluster i and cluster j . Subsequent to clustering, $ESS_{(ij)m}$ is refer to the error sum of square obtained when merging cluster i and cluster j with another cluster, m . The following is the value used

$$ESS_{(ij)m} = \min(ESS_{im}, ESS_{jm}) \quad (5)$$

In this case, $m = 1, \dots, p$ where $m \neq i, j$ and p is number of current cluster. α will be increased since number of cluster become small.

The first rule utilizes the mean and standard deviation to define a significant α . Selection of groups level corresponding to the first stage j , $i = 1, \dots, N-2$ satisfying Equation (6).

$$\alpha_{j+1} > \mu + k\sigma_\alpha \quad (6)$$

where α_{j+1} represents the value of the criterion in stage $j+1$, k is constant, μ and σ_α are respectively the mean and unbiased standard deviation of the α distribution. In this case, k is setup to 1.25 as proposed by Miligan and Cooper (1985).

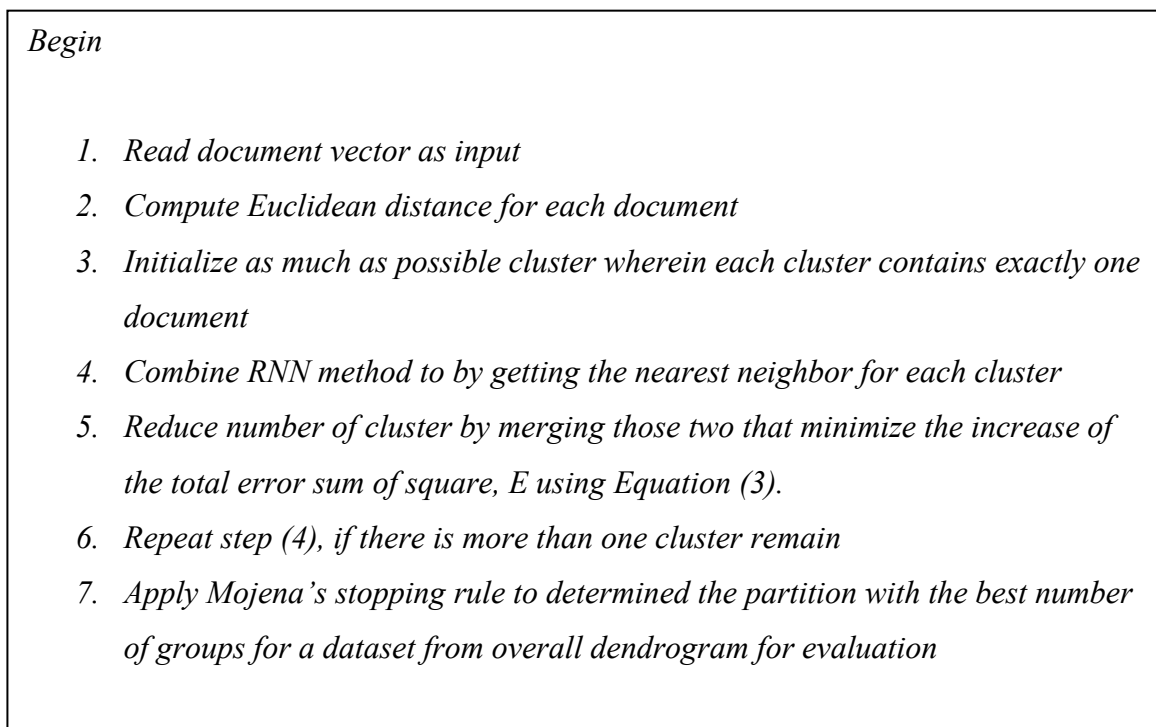


Fig. 3.2 Ward's Algorithm

3.8 Kohonen Clustering

The training set is composed of 206 input variables (title, abstract). A file comprising 206 thesis title and abstract has been used for training and testing phase corresponding to the sample as previously described. 25 interpretation classes characterize the type of supervisor as supervisor code in *Appendix F* and also eight classes for expert as defined in *Appendix G* which list the supervisor code used in the evaluation stage.

3.8.1 PCA implementation

At this stage, the document vector will be reduced into a smaller dimension by considering the meaningful variables. All the data will be processed and eigen values of the meaningful variables will be generated.

Deboeck (1999) explained that Kohonen algorithm result is more effective when used on PCA data compared to the raw data. For this project, Multivariate Statistical Package 3.13m has been used for this purpose. All the 2911 variables produced in section 3.5 were reduced down to 204 variables using PCA further applied to Kohonen network as an input.

The step for reduce dimension using MVSP 3.13m is as follows:

- i. Import document vector into MVSP as an input
- ii. Then select *Analysis* menu and click on *Principal Component*. Principal Component window will then pop-up. Option *Axes to Extract* changed to *All* and maintain others option
- iii. Run the analysis by clicking *Run* button
- iv. The generated output was used for Kohonen clustering

3.8.2 Kohonen Network Algorithm

The following is the Kohonen algorithm due to this project:

- i. Initialize network

Define $w_{ij}(t)$ ($0 \leq i \leq n-i$) to be the weight from input i to node j at time t .

Initialize weights from the n inputs to the nodes to small random values. Set the initial radius of the neighborhood around node j , $N_j(0)$, to be large.

Set the initial cluster based on the labeling process that will be explained in step iii. Learning rate for Kohonen algorithm was setup to 0.005 and three different dimensions will be used in this network. Table 3.2 shows the Kohonen network design for this study.

Table 3.2 Kohonen network design

KOHONEN NETWORK	LearningRate	Dimension	Iteration
	0.005	10x10	5000
5000			7000
12x12		5000	7000
		5000	7000
15x15		5000	7000
		5000	7000

ii. Present input

Present input $x_0(t), x_1(t), x_2(t), \dots, x_{n-1}(t)$, where $x_i(t)$ is the input to node i at time t in terms of document vector after PCA process. A vector is chosen at random from the set of training data and presented to the lattice

iii. Data labeling

Since the author applied the supervised Kohonen, labeling on the training data is needed. The training data is labeled based on the supervisors (considering the supervisor code as in *Appendix F*) for each thesis. At this point, number of clusters depends on the number of different supervisor. In fact, training data is different corresponding to the sample.

iv. Calculate distance using Euclidean function

Compute the distance, d_j between the input and each output node j , given by Equation (7).

$$d = \|X - W_j\| = \left[\sum_{i=1}^n (x_i - w_{ij})^2 \right]^{1/2} \quad (7)$$

where x_i and w_{ij} are the i th elements of the vectors X and W_j , respectively.

v. Select minimum distance

Designate the output node with minimum d_j to be j^* . Every node is examined to calculate which one's weights are most like the input vector. The winning node is commonly known as Best Matching Unit (BMU). The radius of the neighbourhood of the BMU is now calculated. This a value that

starts large, typically set to the ‘radius’ of the lattice but diminishes each time step. Any nodes found within this radius are deemed to be inside the BMU’s neighbourhood.

vi. Update weights

Each neighbouring node’s (the nodes found in step v) weights, $N_{j^*}(t)$ are adjusted to make them more like the input vector. The closer a node is to the BMU, the more its weights get altered. New weights are

$$w_{ij}(t+1) = w_{ij}(t) + \eta(t)(x_i(t) - w_{ij}(t))$$

The term $\eta(t)$ is a gain term ($0 < \eta(t) < 1$) that decreases in time, so slowing the weight adaptation. Notice that the neighbourhood $N_{j^*}(t)$ decreases in size as time goes on, thus localizing the area of maximum activity.

vii. Repeat by going to step 2 for N iteration.

3.9 Evaluation of Ward’s Clustering and Kohonen Clustering Compared to the Expert Survey

Determination of the best technique in this domain study is measure based on the accuracy of percentage corresponding to the expert survey. The accuracy of the entire sample as explained in previous, will be computed and the average percentage for both algorithms will be obtained. The algorithm that produced the highest average accuracy is identified as the best techniques in terms of suggestion of supervisors and examiners.

3.10 Summary

Briefly, this chapter discussed the entire step in order to satisfy the objective of this study. The set of 206 was collected and digitized for preprocessing purposes which are stopword removal and Porter stemming.

In the next stage, the entire document is represented in document vector by considering the *tfidf* weight. The Euclidean distance was selected to measure the distance of the document. Further, Ward's algorithm and Kohonen network is applied to the document vector. The result from Ward's clustering will be compared to the Kohonen result. Here, the correctness percentage for each sample is calculated for evaluation purpose. The next chapter will discuss the results obtained from Ward's and Kohonen network.

CHAPTER IV

RESULT AND ANALYSIS

4.1 Introduction

This chapter discusses and evaluates both clustering techniques based on their performance in suggestion of supervisor and examiner. The result will then be compared with expert survey. The performance of both clustering techniques for suggestion supervisors and examiners based on thesis title and abstract will be evaluated.

4.2 Preprocessing Result

Appendix E shows the preprocessing result for thesis title after the stopword removal and Porter stemming was applied to the 206 set of theses.

4.3 Evaluation of Ward's Clustering and Kohonen Clustering

4.3.1 Ward's Result

Table 4.1 in *Appendix H* show the 15 final clusters produced by Ward's clustering at level 191 corresponding to the Mojena's stopping rule. The unknown thesis had been clustered into several different clusters which contains known documents.

Referring to Table 4.2 in *Appendix H*, Ward's result gives 45.63% accuracy for sample 50:50 when compared to the expert survey.

In the mean time, 82 theses are randomly chosen as testing data in sample 60:40. Only 40 theses from 82 are predicted accurately by Ward's which gives 48.78% accuracy in suggestion supervisor and examiner. The result for 60:40 sample is shown in Table 4.3.

Based on Table 4.4 also in *Appendix H*, 51 theses are identified as a testing data and the rest 165 is identified as a training data for sample 75:25. Ward's produced 45.10% accuracy compared to the expert survey where only 23 theses are predicted accurately.

In particular, Ward's algorithm gives 36.59% accuracy from 41 testing theses that means only 14 theses is predicted accurately for sample 80:20. Detailed result is shown in Table 4.5.

Meanwhile, from 10 testing theses in sample 95:5, 70.00% accuracy is produced by Ward's algorithm as shown in Table 4.6.

Table 4.7 briefly shows the accuracy percentages in determination supervisor and examiner for entire sample involved in this study.

Table 4.7 Ward's Result

Sample	Ward's Result
50:50	45.63%
60:40	48.78%
75:25	45.10%
80:20	36.59%
95:5	70.00%

As we can see in Table 4.7, sample 95:5 shows the highest percentages among the rest sample. This observation Azuaje (2003) results that larger training set will produce more accurate result.

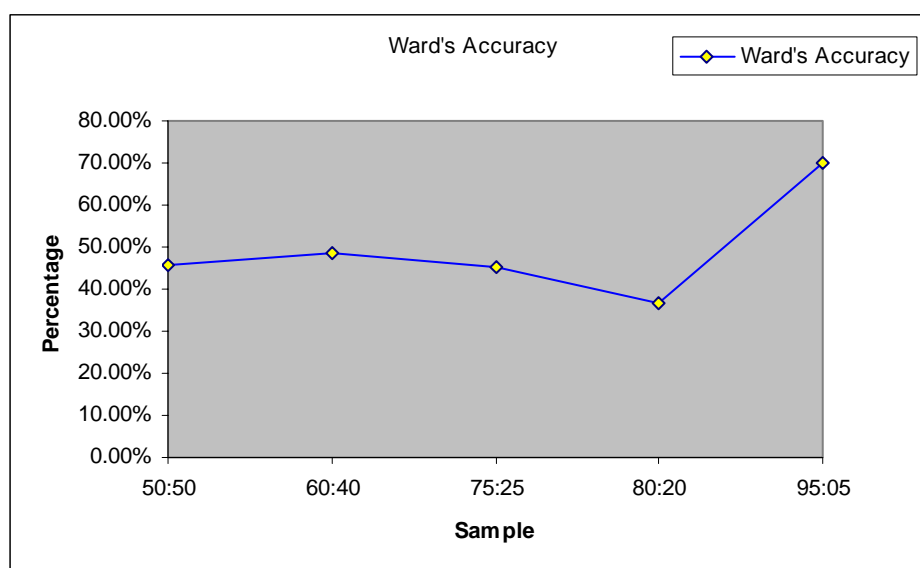


Fig. 4.1 Accuracy of Ward's algorithm

4.3.2 Kohonen Result

There are 204 of variables used in applying Kohonen clustering for substituting 2911 raw data. This 2911 raw data was reduced by using PCA as explained in chapter 3.

Initially, Kohonen clustering used several parameters in order to ensure convergence of the weight. For this thesis, the author had setup three different parameters which are learning rate, iteration and the dimension of the feature map as shown in Table 3.1. As explained in section 3.83, learning rate was setup to 0.005 for each five sample data.

Sample 50:50 was fed to the Kohonen network with 103 testing theses, which producing 43.69 % correctness where only 45 theses is predicted correctly in suggestion of supervisor and examiner. Table 4.8 (*Appendix I*) shows the detailed result for sample 50:50.

Meanwhile, 82 theses was used as testing data in sample 60:40 which gives 42.68% correctness in suggestion of supervisor and examiner. Detailed result can be seen in Table 4.9, *Appendix I*.

Whilst for sample 75:25 in Table 4.10 (*Appendix I*), Kohonen was predicted accurately to the 18 testing theses with 35.26% correctness. In particular, Table 4.11 (*Appendix I*) shows that Kohonen was suggested supervisor and examiner with 46.34% correctness in which 19 theses is predicted correctly compared to the expert survey.

Another last sample, 95:5 shows that Kohonen can suggest supervisor and examiner with 50.00% correctness. This can be referring to Table 4.12 (*Appendix I*).

Table 4.13 Kohonen result

Sample	Kohonen Result
50:50	43.69%
60:40	42.68%
75:25	35.29%
80:20	46.34%
95:5	50.00%

Table 4.13 denoted briefly Kohonen result as explained in previous. Based on this Table 4.13, it shows that sample 95:5 gives the highest accuracy percentage with 50.00%. As discussed in section 4.2.1, larger training set can give better results.

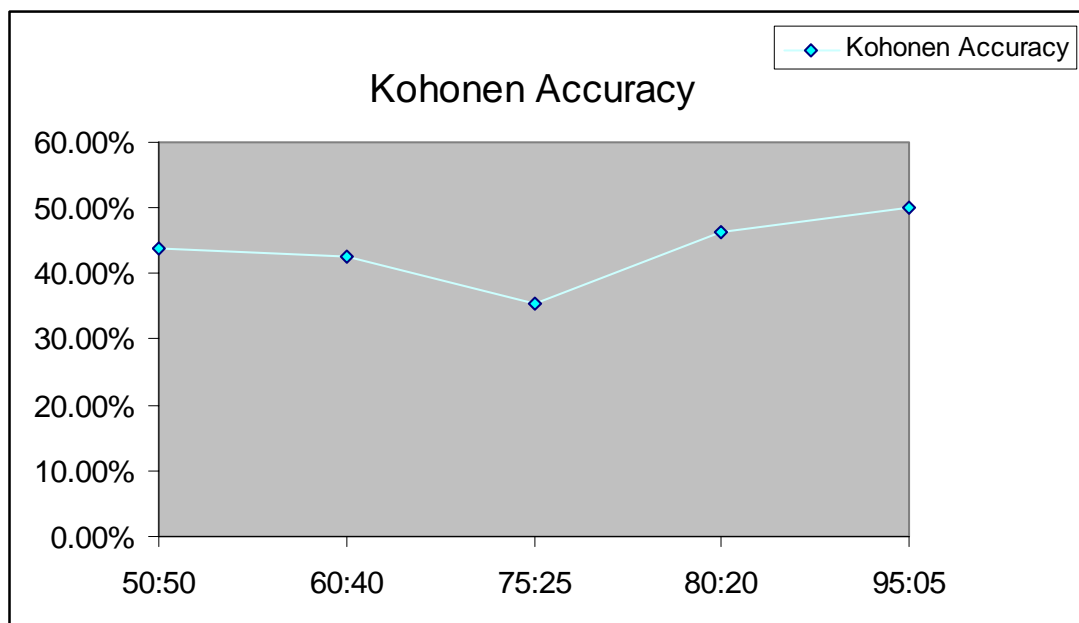


Fig. 4.2 Accuracy of Kohonen algorithm

4.4 Comparative Study and Discussion

Based on the previous section, comparative study has been conducted to find out the best technique for suggestion supervisor and examiner. Table 4.14 shows the comparative study on both clustering techniques.

Table 4.14 Comparative Study

Sample	Ward's Result	Kohonen Result
50:50	45.63%	43.69%
60:40	48.78%	42.68%
75:25	45.10%	35.29%
80:20	36.59%	46.34%

95:5	70.00%	50.00%
AVERAGE	49.22%	43.6%

Considering the clusters result, Ward's shows better performance compared to Kohonen network in suggestion of supervisors and examiners (see Table 4.14). Among five samples, four samples in Ward's algorithm give better performance as highlighted in blue column in Table 4.14. In particular, Kohonen shows best performance only for sample 80:20. Once again, Ward's presents 49.22% accuracy average better than Kohonen network which yielded only 43.6% correctness. Fig. 4.3 shows the performance on both algorithms based on Table 4.14 in suggestion supervisor and examiner.

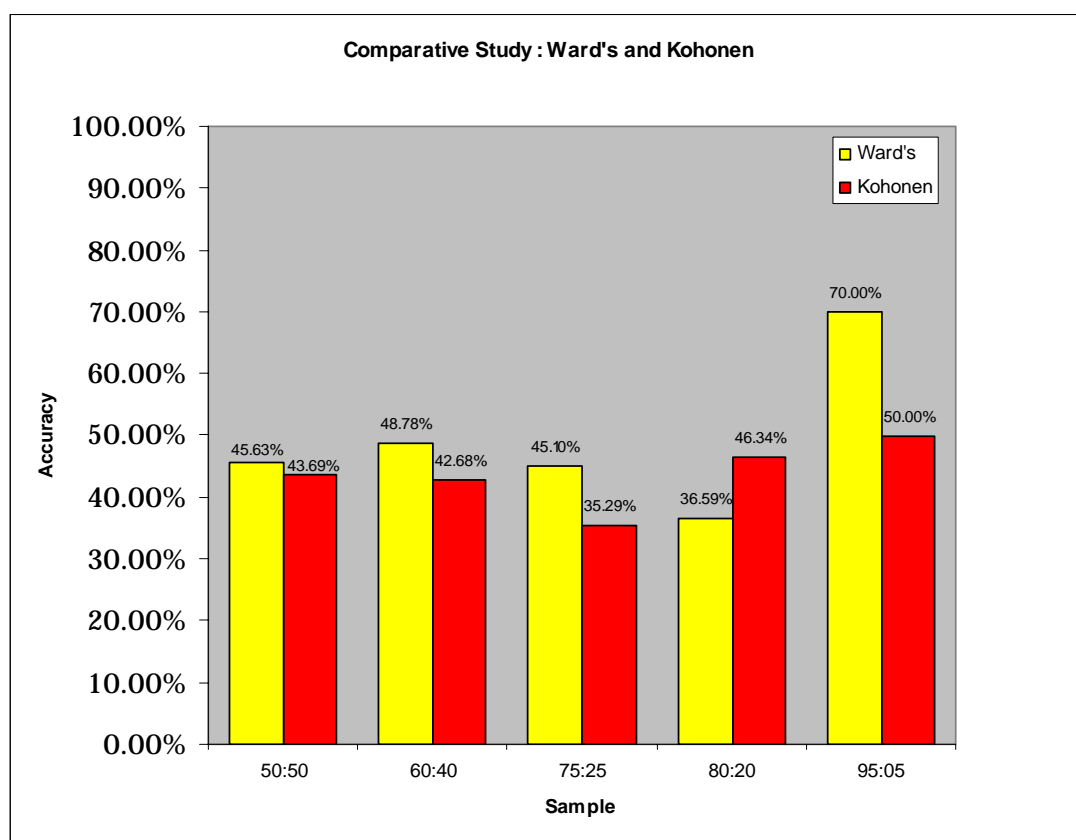


Fig. 4.3 Ward's Clustering vs Kohonen Network

According to the both clustering, better performance is yielded in sample 95:5. As discussed above, the correctness percentage is increased since the training

data set become larger. In fact, the predicted accuracy of a classifier is generally proportional to the size of the training dataset (Azuaje, 2003). This condition is difficult to achieve due to resource and time constraints. More to the point, the size of the training set needs also to be better understood.

However, the increasing of accuracy percentage for Kohonen is in proportion to the increase of available training data. In particular, the possibility of overfitting can also occur in training phase may affect the result produced by Kohonen. It can be seen in sample 75:25 (Fig. 4.2) which the accuracy percentage decreased down to 35.29%.

Since Kohonen performance depends on several parameters such that learning rate, dimension and the number of iteration, and the Kohonen network adaptation has not yet been adequately trialed, it appears a promising technique that warrants further exploration.

Moreover, in order to more accurately evaluate the performance of both algorithms a more thorough analysis is needed. Azuaje (2003) said that accuracy may not tell the whole story for a method. Besides the overall classification success there is a need to evaluate each cluster separately, for precision and recall of specific clusters.

As a conclusion, Ward's presents better performance than Kohonen network in determining supervisors and examiners for FSKSM's Post Graduates theses.

4.5 Summary

Based on the experiments carried out for the analyses, we can conclude that Ward's clustering gives better result when compared to Kohonen network. Actually, the result produced by Ward's in this study is much similar to the result produced by Leuski (2001). It shows that Ward's still maintain quality of cluster and is superior

to Kohonen network. However, the more thorough analysis on both this algorithm is desired to ensure their performance in determining supervisors and examiners.

CHAPTER V

CONCLUSION AND FUTURE RESEARCH

5.1 Summary

The main objective of this study is to measure document clustering performances in determining supervisors and examiners for thesis. From the discussion in literature review, we can see that document clustering has been investigated for use in a number of different areas of text mining and information retrieval. Initially, document clustering was investigated for improving the precision and recall in IR (van Rijsbergen, 1979) and as an efficient way of finding the nearest neighbors of a document (Bukley, et. al, 1985). However, there is only few research in this domain study.

Ward's algorithm and Kohonen network have been adapted in order to evaluate whether document clustering can be used for determining supervisor and examiner. Both algorithms were tested on 206 theses and measured by the percentage of accuracy compared to assignment by human expert. The implementation of Ward's clustering show that Hierarchical clustering has superior performance compares to Kohonen network, though a more thorough analysis is needed to measure both algorithms accurately. It is because the accuracy percentage is presently the initial performance at the same time as precision rather the whole performance.

5.2 Contribution

Based on literature review, the application of document clustering is very limited in this domain study. Subsequently the main contribution of this study is in measuring the performance of document clustering focused on Ward's technique and Kohonen network in suggestion of supervisors and examiners. As expected, Ward's clustering is capable in determining supervisors and examiners effectively. Ward's clustering still maintains its cluster quality compared to Kohonen based clustering in terms of the percentage of accuracy.

Thus, Ward's algorithm can be used in determining supervisors and examiners instead of the manual determination. In addition, human intervention can be avoided since it gives rise to bias in making decision.

Essentially, for further application, we can enhance the used of Ward's clustering techniques for suggestion of project leader for certain research or to find out the best lecturer for certain subjects.

5.3 Further Work

Document clustering is a wide research area and there is still much more things to explore due in the domain. At this point, several suggestions are suggested for future research:

- a) Merging multiple techniques is one of the effective steps; in order to improve IR performance especially for this domain study. This study has only begun to implement many possibilities in term of Ward's clustering and Kohonen clustering in suggestion supervisor and examiner. By merging Ward's and Kohonen techniques, we can attempt to harness the best quality of each technique.
- b) Due to time constraint, only Ward's techniques presented as Hierarchical clustering was applied in this project because its produced quality cluster

(Korenius, et. al. 2004) and for more attestation, another hierarchical clustering techniques should be applied to this domain to find out the best technique in suggestion supervisor and examiner.

- d) Enlarge the data set for better performance because the produced clusters appear inappropriate to each other (Azuaje, 2003). It is because a small test data set may contribute to an inaccurate performance assessment
- e) Apply another clustering technique to this domain to find the best techniques for suggestion supervisor and examiner such as Fuzzy network or non hierarchical clustering techniques

REFERENCES

- Aldenderfer, M. S., & Blashfield, R. K. (1984). *Cluster Analysis*. Sage Publications, Inc.
- Allan, J., Leuski, A., Swan, R., Byrd, D. (2001). Evaluating combinations of ranked lists and visualizations of inter-document similarity. *Information Processing & Management*, 37(3):435-458.
- Anderberg, M.R. (1973). *Cluster Analysis for Applications*. New York: Academic Press.
- Azuaje, H. (2003). *Genomic data Sampling and Its Effect on Classification Performance Assessment*. Northern Ireland, UK.
- Baeza-Yates, R., & Ribeiro-Neto, B. (1999), *Modern Information Retrieval*, ACM Press 1999.
- Borgelt, C., and Nurnbeger, A. *Experiments in Document Clustering using Cluster Specific Term Weights*,
- Botafogo, R.A. (1993). Cluster analysis for hypertext systems. In *Proceedings of the 16th Annual ACM SIGIR Conference*, pp. 116-125. Pittsburgh, PA.
- Chris D. Paice. (1994). An Evaluation Method for Stemming Algorithms. *Proceedings of the Seventeenth Annual International ACM-SIGIR Conference*

Research and Development in Information Retrieval, pages 42-50, 3-6 July 1994

Cormack, R.M. (1971). A review of classification. *Journal of the Royal Statistical Society, Series A*, 134:321-353.

Croft, W.B. (1978). Organizing and searching large files of document descriptions. *Ph.D. Thesis*, Churchill College, University of Cambridge.

Cutting, D.R., Karger, D.R., Pedersen, J.O., Tukey, J.W. (1992). Scatter/Gather: A cluster based approach to browsing large document collections. In *Proceedings of the 15th Annual ACM SIGIR Conference*, pp. 126-135. Copenhagen, Denmark.

Defays, D. (1977). An efficient algorithm for a complete link method. *Computer Journal*, 20:93-95.

Dawson, J.L. (1974): "Suffix removal for word conflation," *Bulletin of the Association for Literary & Linguistic Computing*, 2 (3), 33-46.

Doszkocs, T. E., Reggia, J., & Lin, X. (1990). Connectionist models and information retrieval. *Annual Review of Information Science and Technology (ARIST)*, 25, 209-260.

Dunlop, M. D. Development and evaluation of clustering techniques for finding people, *Proc. of the Third Int. Conf. on Practical Aspects of Knowledge Management (PAKM2000) Basel, Switzerland*, 30-31 Oct. 2000,

El-Hamdouchi, A. and Willett, P. (1989). Comparison of hierarchic agglomerative clustering methods for document retrieval. *The Computer Journal*, 32(3):220-227.

Griffiths, A., Robinson, L.A., Willett, P. (1984). Hierarchic agglomerative clustering

methods for automatic document classification. *Journal of Documentation*, 40(3):175-205.

Griffiths, A., Luckhurst, C., and Willett, P., "Using interdocument similarity information in document retrieval systems. *Journal of the American Society for Information Science*, 37:3-11.

Harman, D. (1991). How effective is suffixing? *Journal of the American Society for Information Science*, 42, 1(1991), 7-15.

Hartigan, J.A. (1975). *Clustering algorithms*. New York: Wiley.

Hearst, M.A. and Pedersen, J.O. (1996). Re-examining the Cluster Hypothesis: Scatter/Gather on Retrieval Results. In *Proceedings of the 19th Annual ACM SIGIR Conference*, pp. 76-84. Zurich, Switzerland.

Hideo Fuji, W. Bruce Croft, *A Comparison of Indexing Techniques for Japanese Text Retrieval*,

Honkela, T., Kaski, S., Lagus, K., and Kohonen, T. (1996) "Newsgroup exploration with WEBSOM method and browsing interface". Report A32, Faculty of Information Technology, Helsinki University of Technology (Rakentajanaukio 2 C, SF-02150 Espoo, Finland).

Jardine, N. and Sibson, R. (1968). The construction of hierarchic and non-hierarchic classifications. *Computer Journal*, 11(2):177-184.

Jardine, N. and van Rijsbergen, C.J. (1971). The use of hierarchical clustering in information retrieval. *Information Storage and Retrieval*, 7:217-240.

Kandel, A., Schenker, A., Last, M. and H. Bunke, (2003). A Comparison of Two Novel Algorithms for Clustering Web Documents, *Proceedings of the 2nd International Workshop on Web Document Analysis (WDA 2003)*, pp. 71-74, Edinburgh, Scotland.

- Kohonen, T. (1997). *Self-Organizing Maps*. 2nd ed., Springer-Verlag, Berlin.
- Korenius, T., Laurikkala, J., Jarvelin, K., Juhola, M. (2004). Stemming and Lemmatization Finnish document. *ACM Conference on Information and Knowledge Management (CIKM)*
- Krovetz, R., et. al., Viewing morphology as an inference process, *Proc. 16th ACM SIGIR Conference*, Pittsburgh, June 27-July 1, 1993; pp. 191-202.
- Lancaster, F. W. (1979). *Information retrieval systems : characteristics, testing and evaluation*, 2nd ed., New York, John Wiley.
- Lance, G.N. and Williams, W.T. (1967). A general theory of classificatory sorting strategies. I. Hierarchical systems. *Computer Journal*, 9:373-380.
- Leuski, A. (2001). Interactive information organization: techniques and evaluation. *Ph.D. Thesis*, University of Massachusetts, Amherst.
- Leuski, A. (2001). Evaluating document clustering for interactive information retrieval. *Proceedings of Tenth International Conference on Information and Knowledge Management (CIKM'01)*, pages 41-48, Atlanta, Georgia, USA,
- Lin, X., Soergel, D.. & Marchionini, G. (1991. October). A self-organizing semantic map for information retrieval. In *Proceedings of the Fourteenth Annual International ACM/SIGIR Conference on Research und Development in I&rmation Retrieval* (pp. 262-269). Chicago, IL.
- Lovins J.B., 1968: "Development of a stemming algorithm," *Mechanical Translation and Computational Linguistics* **11**, 22-31.
- Macskassy, S.A., Banerjee, A., Davidson, B.D., Hirsh, H. (1998). Human

performance on clustering web pages: a preliminary study. In *Proceedings of The 4th International Conference on Knowledge Discovery and Data Mining (KDD-98)*, pp. 264-268. New York, NY.

Maron, M. E., & Kuhns, J. L. (1960). On relevance, probabilistic indexing and information retrieval. *Journal of the ACM*, 7, 2 16-243.

Milligan, G. W. and Cper, M.C. (1985). An Examinatin of Procedures for Determining the Number f Clusters in a Data Set. *Psychometrika*. 50: 159-179

Milligan, G.W., Soon, S.C., Sokol, L.M. (1983). The effect of cluster size, dimensionality, and the number of cluster on recovery of true cluster structure. *IEEE Transactions on Patter Recognition and Machine Intelligence*, 5(1):40-47.

Mirkin, B. (1996). *Mathematical Classification and clustering*. Kluwer

Mock, K. (1998). A Comparison of Three Document Clustering Algorithms: TreeCluster, Word Intersection GQF, and Word Intersection Hierarchical Agglomerative Clustering. *Intel Technical Report*

Mojena, R. (1977). Hierarchical Grouping Methods and Stopping Rules: An Evaluation. *The Computer Journal*, 20: 359-363

Murtagh, F. (1985). Multidimensional Clustering Algorithm, *COMPSTAT Lectues 4*, *Physic-Verlag*, Vienna

Na Tang and Rao Vemuri, V. (2004). Web-based Knowledge Acquisition to Impute Missing Values for Classification, *IEEE/WIC/ACM International Joint Conference on Web Intelligence*, Beijing, China

Pirkola, A. (2001). Morphological typology of languages for information retrieval. *Journal of Documentation*, 57, 3 (2001), 330-348.

- Porter, M.F. (1980). An Algorithm for Suffix Stripping. *Program - Automated Library and Information Systems*, 14(3): 130-137.
- Qin Ding et al., *Data Mining Survey*, North Dakota State University.
- Qin He, (1999). *Neural Network and Its Application*, Spring, University of Illinois at Urbana-Champaign
- Reggia, J. A.; & Sutton, G. G., III. (1988). Self-processing networks and their biomedical implications. *Processings of the IEEE*, 76, 680-692.
- Salton, G. and Wong, A. (1978). Generation and search of clustered files. *ACM Transactions on Database Systems*, 3(4):321-346.
- Salton, G. and McGill, M.J. (1983). *Introduction to Modern Information Retrieval*. New York: McGraw-Hill.
- Salton, G. and Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing & Management*, 24:513-523.
- Sibson, R. (1973). SLINK: an optimally efficient algorithm for the single link cluster method. *Computer Journal*, 16:30-34.
- Sneath, P.H.A. and Sokal, R.R. (1973). *Numerical taxonomy: the principles and practice of numerical classification*. San Francisco: W.H. Freeman.
- Sparck Jones, K., (1972). "A statistical interpretation of term specificity and its application in retrieval", *Journal of Documentation*, v28, pp 11-21, 1972.
- Steinbach, M., Karypis, G., and Kumar, V. (2000). A comparison of document clustering techniques. *KDD Workshop on Text Mining*. D.C., USA
- Sudipto, G., Rajeev, R., and Kyuseok, S. (1998). CURE: An efficient clustering algorithm for large databases. In *Proc. of 1998 ACM SIGMOD Int. Conf. on Management of Data*, 1998.

- Sudipto, G., Rajeev, R., and Kyuseok, S. (1999). ROCK: a robust clustering algorithm for categorical attributes. In *Proc. of the 15th Int'l Conf. on Data Eng.*, 1999.
- Turtle, H.R. and Croft, W.B. Evaluation of an inference network-based retrieval model. *ACM Trans. Inf. Syst.* **3** (1991), 187-222.
- van Rijsbergen, C.J. (1971). An algorithm for information structuring and retrieval. *Computer Journal*, 14:407-412.
- van Rijsbergen, C.J. and Sparck Jones, K. (1973). A test for the separation of relevant and non relevant documents in experimental retrieval collections. *Journal of Documentation*, 29(3):251-257.
- van Rijsbergen, C.J. and Croft, W.B. (1975). Document clustering: An evaluation of some experiments with the Cranfield 1400 Collection. *Information Processing & Management*, 11:171-182.
- van Rijsbergen, C.J. (1979). *Information Retrieval*. London: Butterworths, 2nd Edition.
- Voorhees, E.M. (1985a). The effectiveness and efficiency of agglomerative hierarchic clustering in document retrieval. Ph.D. Thesis, Technical Report TR 85-705 of the Department of Computing Science, Cornell University.
- Voorhees, E.M. (1986). Implementing Agglomerative Hierarchic Clustering Algorithms for Use in Document Retrieval. *Information Processing & Management*, 22(6): 465-476.
- Ward, J.H. (1963). Hierarchical grouping to minimize an objective function. *Journal of the American Statistical Association*, 58:236-244.
- Weiss, R., Velez, B., Sheldon, M. (1996). HyPursuit: A hierarchical network search

engine that exploits content-link hypertext clustering. In *Proceedings of Hypertext '96*, pp. 180-193. Washington, DC.

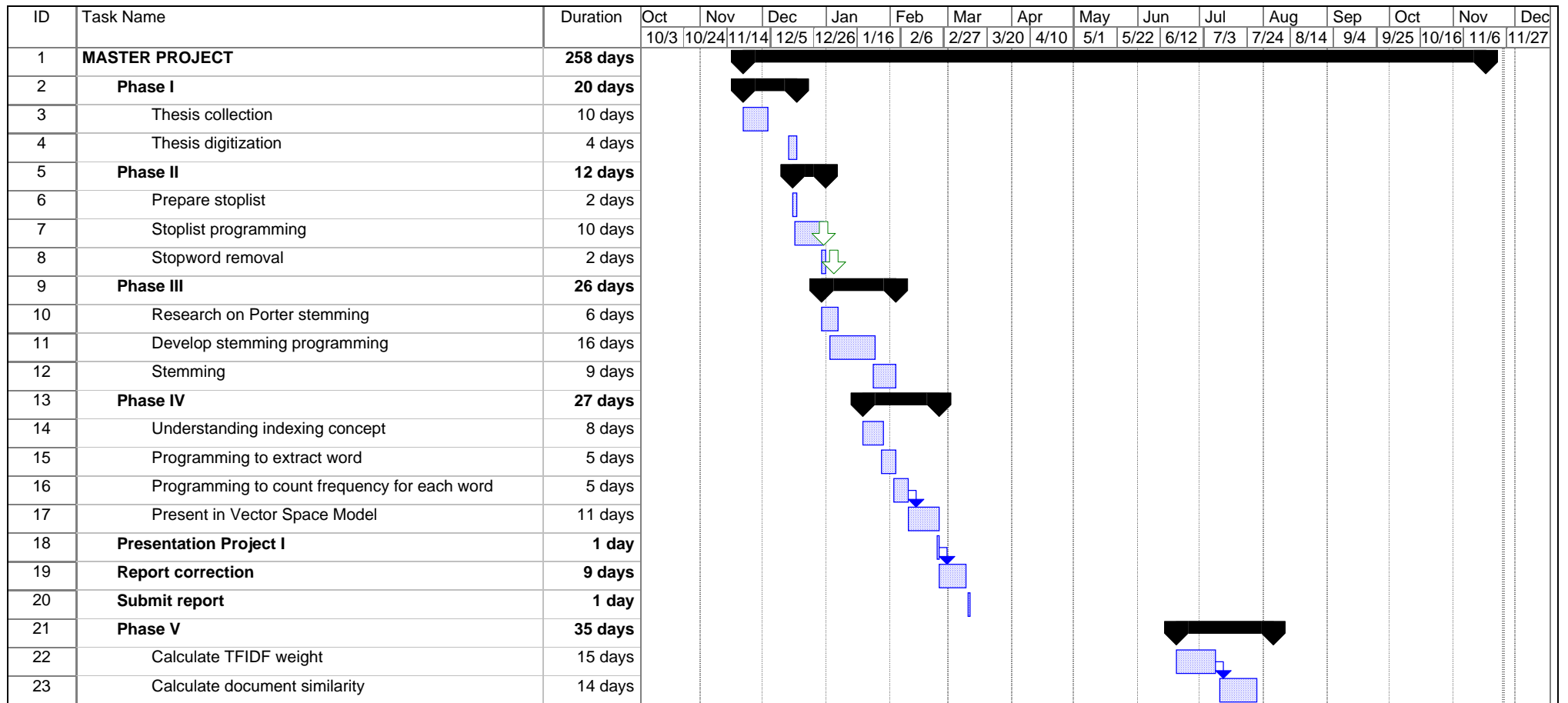
Williams, W.T., Clifford, H.T., Lance, G.T. (1971a). Group size dependence: a rationale for choice between numerical classifications. *Computer Journal*, 14:157-162.

Willett, P. (1988). Recent trends in hierarchic document clustering: A critical review. *Information Processing & Management*, 24(5):577-597.

Wishart, D. (1969). An algorithm for hierarchical classification. *Biometrics*, 25:165-170.

Zamir, O. and Etzioni, O. (1988). Web document clustering: A feasibility demonstration. In *Proceedings of the 21st Annual ACM SIGIR Conference*, pp. 46-54. Melbourne, Australia.

APPENDIX A
Gantt Chart



Project: gantt chat Date: Fri 11/25/05	Task		Milestone		External Tasks	
	Split		Summary		External Milestone	
	Progress		Project Summary		Deadline	

APPENDIX B
Thesis Collection

No	Title	Supervisor	ExpertSurvey
1	A Comparative Study Between Smart School And Normal School On The Usage Of Ict In Johor Bahru	Prof. Dr. Rose Alinda Binti Alias	5
2	A Decision Support System Based On Neural Networks To Select Candidates For The Computer Science Degree Program	Prof.Madya Dr.Mohd Noor Bin Md Sap	1
3	A Prototype Virtual Intelligence Map System	Prof Dr Ahmad Zaki b Abu Bakar	5
4	Resource usage analyzing for distributed threats simulation on intrusion detection system hose	PM Dr Mohd Azaini b Maarof	2
5	Active Reaction analyzing for distributed threats simulation in parallel on Intrusion Response system	PM Dr Mohd. Aizaini Maarof	2
6	Knowledge Management Application Support Towards the Lecturer Communities of Practice In Institute Of Higher Learning	Prof. Madya Dr. Shamsul bin Shahibuddin, En. Mohammad Nazir bin Ahmad @ Sharif	5
7	Application of Fuzzy Logic on Decision Support Systems For Selecting Subject Course In Universiti	PM. Abd. Manan Ahmad	1
8	Class Timetabling Using Modified Genetic Algorithm	PROF. MADYA DR. SAFAAI BIN DERIS	1
9	Computer Controlling System	P.M. Dr. Mohd Aizaini Bin Maarof	2
10	Content Management Framework For Malaysian Government Web Sites	PROF. ZAMRI B. MOHAMED	5
11	Critical Success Factors For Managing Dot.Com Company	PM DR HARIHODIN B SELAMAT	5
12	Denial Of Service Attack Detection	Prof Dr Abdul Hanan b Abdullah	2
13	Developing A University Knowledge Community Using Sms Technology	PM Dr. Rose Alinda Alias, Professor Dr. Ahmad Zaki bin Abu Bakar	5
14	Developing A Web Based Tourism Information System For Sarawak Tourism Board	Prof Zamri b Mohamed	5
15	Development of A Prototype For Johore Tourism Information System In Web Environment	PM Dr. Mohd. Noor bin Md. Sap	5
16	Easy Link Information Centre Administrator (Elica)	PM Dr Shamsul b Sahibuddin	5
17	Electronic Commerce For A Computer Shop	PM Dr Harihodin b Selamat	7
18	Enhancing Customer Relationship Process In Southern Sumatra District Office Of P.T. Telekomunikasi Indonesia, Tbk. Through Integration Of Multiple Communication Services	PM Dr Abdul Samad b Hj Ismai	4
19	Guide Line for Information and Communication Technology Management in Small Medium Industries Countryside	PROFESSOR ZAMRI BIN MOHAMED, Encik Md. Hafiz Bin Selamat, Encik Mohd. Zaidi Bin Abdul Rozan	5
20	Beam Search Implementation in Solving Personal Computer Configuration Problem	PM Abdul Manan Ahmad	1
21	Implementation Of Constraint Based In Scheduling Nurse Shift At Crystal Ward, Hospital University Malaysia	Professor Dr. Safaai Bin Deris, PM Safie Bin Mat Yatim, PM Abdul Manan	1
22	Information Security Policy For Universiti Teknologi MARA	Prof Dr Abdul Hanan b Abdullah	2
23	Integration of Workspace Awareness in Collaborative Case Tools	PM DR. Shamsul Sahibuddin, Pn Mazleena	4
24	Islamic E-Organizer	PM Dr Shamsul b Sahibuddin	5
25	Guide Line Forming to Improve Effectiveness of Supervising and Controlling ICT Project at government agency by Project Leader Committee	Prof Zamri b Mohamed	7
26	Reengineering Magazines Ordering System : Replacement Structured System and Analysis to Object Oriented Design (UML)	PM Dr Mohd Noor b Md Sap	1
27	Knowledge Cents As A Contribution Metrics For Knowledge Management	Prof Dr Ahmad Zaki b Abu Bakar, En Mohd Zaidi b Abdul Rozan	5
28	Managing A Tutoring System For Propagating Plants-Living Skills Subject	PM Dr Rahman b Ahmad	8
29	Mobile Protection System	Assoc. Prof. Dr. Shamsul Bin Sahibuddin	4
30	Neural-Fuzzy For Handwritten Digits Recognition	PM Dr Siti Mariyam bt Hj Shamsudin	3
31	Development of Computer Games Engine on simulation strategy style	Prof. Dr. Ahmad Zaki Abu Bakar	5
32	FSKSM Sharenet Implementation in Encourage Knowledge Sharing Process	PM Abdul Manan b Ahmad	1
33	System Prototype for Tools Technics Planning Information System	PM Dr Rose Alinda Alias	7
34	Weight Using Method Implementation for Strategic Management Assessment System for IPTS : Review Case in ITP-YPJ	PM DR ROSE ALINDA BT ALIAS, Pn Zeti Darleena	5
35	An Object Oriented Prototype System For A Small And Medium Enterprise	PM Dr Rose Alinda bt Alias	5
36	Implementation Infomediary Concept inE-Learning Environment	PM Dr Rahman b Ahmad	8
37	Hypermedia Application Model using Combination UML and HDM	Prof Dr Safaai b Deris	1
38	Recurrent Neural Network in Prediction House Price	PM Dr Siti Mariyam bt Hj Shamsudin	1
39	Enhancement MOO Tool in Distance Learning	PM Dr. Shamsul Sahibuddin	7
40	Effective Hybrid Bounding Volume Serial Strategic for Detecting Collision N-Rigid Convex Object in an Interactive Virtual Environmen	PM Daut Daman	3

41	Speech Recognition For Isolated Word Samples Using Support Vector Machine	PM DR SITI MARIAM BINTI HJ SHAMSUDIN, EN MD SAH BIN HJ SALAM	3
42	Gray Scale Fingerprint Image Minutiae Detection Using Ridge Line Following Algorithm	Prof Dr Ghazali Sulong	3
43	Neural Network in time series forecasting	PM Dr Salihin b Ngadiman	1
44	Implementation of Knowledge Management in Higher Learning Institute: Case study FSKSM	PM Dr Shamsul b Sahibuddin, Pn Norhawaniah bt Zakaria	5
45	Rainfall Forecasting Using Neural Network based on Massive Meteorological Data	P.M. Dr. Mohd Noor bin Md. Sap	1
46	Comparison of Classification Result for Undiscretized and Discretized Data : Back Propagation and Rough Set	PM Dr Siti Mariyam bt Hj Shamsudin	1
47	Neural Network Optimization using Genetic Algorithm in Speech recognition	PM Dr Zulkifli Mohamad	3
48	Enhancement Dick, Carey and Carey (2001) Model using A Nine Step Model in developing intelligence learning system	PM Noraniah bt Mohd Yassin	8
49	Information system of Quality Service Assessment :Customer Automated Support System (CASS)	PROF MADYA DR ROSE ALINDA BTE ALIAS	5
50	Comparison of linear summation technique and neural network model in decision support for student intake in Higher Education Institution	PM Dr Mohd Noor b Md Sap	1
51	Performance Comparison of Java RMI and CORBA in Multi-Level Marketing Business Structure	PM. DR. Rose Alinda Alias	5
52	Collaborative environment for JPA top level management : Management and Professional	PM Dr Shamsul b Sahibuddin	5
53	Properties Consultation Computerized	Pm Abdul Manan b Ahmad	1
54	Knowledge Audit Portal for Public Higher Education	PM Dr Rose Alinda Alias	5
55	Small Medium Industries Community Portal	PM Safie Mat Yatim, En Norhashim Abu Samah	5
56	UPSI Education Community Portal	PM Dr Shamsul Sahibuddin	5
57	Knowledge Management System	Prof Dr Ahmad Zaki b Abu Bakar	5
58	Seminar Management and Monitoring Portal	Prof Dr Safaai b Deris	1
59	Product Information Searching Through WAP	PM Dr Mohd Azaini b Maarof	2
60	Project Time Management And Communication System	Prof Dr Safaai b Deris	1
61	Protein Secondary Structure Prediction From Amino Acid Sequences Using A Neural Network Classifier Based On The Dempster-Shafer Theory	Prof Dr Safaai b Deris, PM Dr Rosli Md Illias	1
62	Evaluation System for E-Learning Portal	PM Dr Mohd Noor b Md Sap	5
63	World Islamic Trade Business Game (WITNES) in Prototype	En Noor Azam b Mohd Sheriff	3
64	Prototype System of Personal Firewall	Prof. Dr. Abdul Hanan Bin Abdullah	2
65	Prototyping A Profit And Loss Statement Analysis Using Simulation Modeling And Fuzzy Logic	Prof Dr Ahmad Zaki b Abu Bakar	5
66	Region Based Digital Image Segmentation Using Pixel Caste-Mark And Pixel Discrimination	PM Dr Mohd Noor b Md Sap, Pm Dr Harihodin b Selamat	1
67	Tools Design Web Based In Supporting Collaborative Learning	PM Dr Harihodin b Selamat	5
68	Trap detached design in Simple Network Management Protocol (SNMP) using wireless application protocol (WAP) Client Server System	Prof Dr. Shamsul Sahibuddin	4
69	Intelligent Mobile Agent Open Architecture for Distributed Application	Prof Dr Safaai b Deris	1
70	Decision Support System using Neural Network in Bank Loan Application	PM Dr Mohd Noor Md Sap	1
71	School Discipline Decision Support System	PM Abdul Manan b Ahmad	1
72	decision support system of service quality information system in IPTS	PM Dr Rose Alinda bt Alias	5
73	Agency Management Smart System	PM. SAFIE MAT YATIM, PM. DR. SAFAAI DERIS	5
74	Quality Assurance System at Top Empire Sdn Bhd	Prof. Madya Dr. Ab. Rahman Ahmad	6
75	Software Standardized Control System	Prof Dr. Hanan b Abdullah	2
76	Assembly Line Balancing Workstation System Using Heuristic Method	Prof. Madya Dr. Ab Rahman Ahmad, Dr. Masine bte Md Tap	3
77	Student Information System	En Nadzari b Shaari, Prof Dr Safaai b Deris	5
78	student discipline system in school involving merit demerit process	PM Abdul Manan b Ahmad	1
79	Electronic Document Delivery System	PM Dr Mohd Noor b Md Sap	5
80	Data Recovery System in Disc Forensic for Windows Operating System	PROF MADYA DR. SHAMSUL SAHIBUDDIN	5
81	Intelligent tutoring system : Possibilities Statistic Topic	PM Noraniah bt Mohd Yassin	8
82	Faraid knowledge information Based on Web technology	PM Abdul Manan Ahmad	5
83	Staff Information Management system Based on Lotus Notes : Case Study at Mydin Mohammad & Sons Sdn Bhd	Dr Shamsul b Sahibuddin	5
84	Knowledge Management System in Solving ISO 9000	Dr. Azizah Binti Abdul Rahman, En. Azlan Bin Mohd Zain	5
85	Student Performance Assessment System	PM Noraniah Mohd.Yassin	8
86	Online Vehicle Sale System	P.M. Safie Mat Yatim, P.M. Dr.	5

		Mohd. Noor Md. Sap	
87	Development for the Income Statement and Individual Income Tax Calculation On-Line System	PM Dr Shamsul b Sahibuddin	5
88	Customer Profile System (CPS) : Case Study at Telekom Malaysia Berhad	Prof. Madya Dr. Harihodin Selamat	7
89	Decision Support System for Vehicle Buying Planning through Hire Purchase	PM Dr Shamsul b Sahibuddin, Pn Norhawaniah bt Hj Zakaria	5
90	The Application Of Enhanced Genetic Algorithm In Class Timetabling Problem	Prof. Dr. Safaai bin Deris	1
91	The Effect of Malaysian Smart School To Public Universities Curriculum Structures In Term Of Basic IT Subjects	PM Abdul Manan b Ahmad	1
92	The Retrace Traveling Salesman Problem: A New Extension Of Traveling Salesman Problem	PROF. MADYA DR. SAFAAI BIN DERIS	1
93	Trademark Matching Algorithm based On Simplified Feature Extraction	Dr Dzulkifli Mohamad	3
94	Visualization for Large Data Set Of Triangular DTM	PM Daut Daman	3
95	Business Advertisement via Internet	PM Dr Mohd Azaini b Maarof	2
96	Workflow Management System for Strata Title Application At Federal Lands And Mines Department	PM Dr Rose Alinda bt Alias	5
97	System Analysis Comparison between Data Flow Diagram and Use Case	PM Dr Rose Alinda Binti Alias	5
98	The Recognition of Plate Number Location Using Statistical Method	Dr. Dzulkifli Mohamad	3
99	Transition of the System Design from a Functional Oriented To an Object Oriented	Pn Azizah binti Abdul Rahman, Pn Nor Hawaniah binti Zakaria	5
100	The Effect of Architecture on Exact Forecasting in Backpropagation	Prof Dr. Safaai bin Deris	1
101	Electronic Commercial System based on Letter Credit: Case Study at PT. Khage Lestari Timber	PM. Dr. Abdul Hannan Abdullah	2
102	Building a Prototype Data Warehouse a Case Study at FAMA	PM Dr. Abdul Hannan Abdullah	2
103	Leadership Through Knowledge Management Portal: A Prototype	Prof. Dr. Ahmad Zaki bin Abu Bakar	5
104	Exploring the Notion of Service Assessment within the Context of Information System Service Quality (ISSQ) in the Malaysian Public Service	Associate Professor Dr. Rose Alinda binti Alias	5
105	Decision Support System for Rural Digital Divide Programs	Prof. Zamri bin Mohamed	7
106	Development of Web Creation and Management Tool for Education Purposes Based On Web	PM Noraniah Mohd Yassin	8
107	Developing MSCIT E-Learning Portal Prototype: Case Study of FSKSM MSc. IT Programme	Dr. Azizah binti Abdul Rahman	5
108	A Prototype of Flood Management Support System National Flood Forecasting Center, Department Irrigation and Drainage, Malaysia	PM Dr. Rose Alinda binti Alias	5
109	Enhancing Decision Making Processes in Project Monitoring Environment in Planning and Development Division, Ministry of Health	Professor Zamri bin Mohamed	7
110	The Management Information System for Overall Equipment Efficiency (OEE) Analysis and Decision Making	PM Dr. Safaai bin Deris	1
111	Health Promotion E-Portal	PM Dr. Shamsul bin Shahibuddin	5
112	Information System Plan for Secondary School: Case Study at SMK Mutiara Rini	PM Dr. Rose Alinda binti Alias	7
113	Decision Support System for Personal Budget	PM Dr. Shamsul bin Sahibuddin	5
114	Electronic Claim Management System	PM Abdul Manan Ahmad	1
115	Computer Assisted Learning-Algebra Fraction using EIF approach	PM Dr. Ab. Rahman bin Ahmad	8
116	Pornographic Web Page Filtering System Using Neural Network Model	PM Dr. Siti Mariyam Hj. Shamsuddin	1
117	Zakat Payment System for Jabatan Agama Islam Johor	PM Dr. Safaai bin Deris	1
118	A Mobile and Wireless Ward in Hand System: (WiH)	PM Dr Rose Alinda bin Alias	5
119	School Management Information System	PM Safie bin Mat Yatim	8
120	Smart Kindergarten Management System	Dr. Muhammad Shafie Hj Abdul Latiff	4
121	Instituting Knowledge Sharing Among Senior Officers in the Prison Department of Malaysia	Professor Zamri bin Mohamed	7
122	Time Table Scheduling System for Primary School	PM Abdul Manan Ahmad	1
123	Activity Based Costing Software for Manufacturing Industries	Dr. Muhammad Shafie Abdul Latiff	4
124	Knowledge Classification System for Sistem Saraan Malaysia	Prof. Dr. Ahmad Zaki bin Abu Bakar	5
125	Online Data Storage : Drivepods	PM Dr. Rose Alindabinti Alias	5
126	Business Development Automation Using Market Basket Analysis Techniques	PM Abdul Manan Ahmad	1
127	Academic Advisor Expert System	PM Dr. Safaai bin Deris	1
128	Teacher Performance Appraisal System	PM Dr. Mohd Noor bin Md. Sap	1
129	Acquisition Online System: Case Study In PSZ	PM Dr. Mohd Noor bin Md. Sap	1
130	Designing and Implementing Double Cube Data Model	PM Dr. Harihodin bin Selamat, PM Daut Daman	7
131	Information Retrieval System Using Text Block Indexing	PM Safie bin Mat Yatim, PM Sarudin bin Bakri	3
132	Information Security Policy Of Jabatan Kastam Diraja Malaysia, Johor	PM Dr. Mohd. Aizaini Maarof	2

133	Cluster-Based Compound Selection Using Fuzzy Clustering	PM Dr. Naomie binti Salim, Dr Ali bin Selamat	1
134	Online Journal Management System for Journal Information Technology	PM Dr. Naomie bin Salim	1
135	Cluster Analysis on Chemical Data using Genetic Algorithm	PM Dr. Naomie binti Salim, Dr. Ali bin Selamat	1
136	Task Monitoring and Productivity Management System (Case Study: SPMB Workshop)	Dr. Azizah binti Abdul Rahman	5
137	Decision Support System Assigning Machine's at Top Empire Industries Sdn. Bhd	PM Dr. Abd. Rahman bin Ahmad	6
138	Web Based Job Application System	PM Dr. Shamsul Sahibuddin	5
139	Prediction of Life Expectancy for Patients with Hepatitis Using Support Vector Machines and Wrapper Method	Professor Dr. Safaai bin Deris	1
140	Implementation of Lot Sizing and Forward Wagner Whitin Method in Rolling Horizon Environment	PM Dr. Mohd Salihin bin Ngadiman	1
141	Assessment Performance Candidate Finding System	Prof. Dr. Ahmad Zaki bin Abu Bakar	8
142	Using Genetic Algorithm with Directed Mutation in Solving Timetabling Problems	Assoc. Prof. Dr. Mohd Salihin bin Ngadiman, Puan Roselina binti Sallehudin	1
143	Human Animation using Neural Network	PM Dr. Siti Mariyam binti Shamsudin	1
144	Comparison of the Effectiveness of Probability Model with Vector Space Model for Compound Similarity Searching	PM Dr. Naomie binti Salim, Puan Razana Alwee	1
145	Measurement System Analysis (MSA) in Automotive Manufacturing Industry using GR & R	PM Dr. Mohd Salihin bin Ngadiman	6
146	Bioactivity Classification of Anti AIDS Compounds Using Neural Network and Support Vector Machine: A Comparison	Assoc. Prof. Dr. Naomie binti Salim	1
147	Identification of Bioactivity Molecule for AIDS : A Comparison of Neural Network and Rough Set	PM Dr. Naomie binti Salim	1
148	Finding Best Coefficient and Fusion of Coefficients for Similarity Searching Using Neural Network Algorithms	PM Dr. Naomie binti Salim	1
149	Pairwise sequence alignment for selection of effective substitution matrices and gap penalty parameter for sequence alignment in dynamic Programming	PM Dr. Naomie binti Salim, Encik Muhamad Razib bin Othman	1
150	Promoting Reflective Practice in UTM's Teaching Community with the User of Information Technology (IT)	PM Dr. Rose Alinda binti Alias, PM Dr. Abdul Samad bin Ismail	5
151	Prototype of an e- Learning assessment application based on Bloom Taxonomy for Physics Form 4	PM Dr. Mohd Noor bin Md. Sap	8
152	Knowledge Management System for Managing Rosettanet Implementation in Johore	Prof. Dr. Ahmad Zaki bin Abu Bakar, En. Md. Hafiz bin Selamat	5
153	Rain Distribution Forecasting By Clustering In Data Mining: A Comparison of Association Rules Technique and Statistical Method	PM Dr. Mohd. Noor bin Md. Sap	1
154	A Study on Entrepreneurial Intention among Information technology Technopreneurs	Prof. Dr. Ahmad Zaki bin Abu Bakar	5
155	Features Extraction For Protein Homology Detection Using Hidden Markov Models Combining Scores	Nazar M. Zaki, Safaai Deris , Rosli M. Illias	1
156	Electrical Appliances Control System Via Internet Based On Parallel Port	Prof Dr. Abdul Hanan bin Abdullah	2
157	Development of Surface Reconstruction For Ship Hull Design	Fadni Bin Forkan, Mahmoud Ali Ahmed, Ang Swee Wen, Siti Mariyam Hj. Shamsuddin, Cik Suhaimi Bin Yusof, Mohd. Razak Samingan, Yahya Samian	3
158	CSCW System In Office Environment Application	Prof Dr Mohd Aizaini Maarof	2
159	Fingerprint Classification Approaches: An Overview	Leong Chung Ern, Dr. Ghazali Sulong	3
160	An Hybrid Trust Management Model For MAS Based Trading Society	Prof Dr. Aizaini Maarof, Krishna K.2	2
161	Interaction between Agents (Arguing and Cooperating Agents)	Ng Kee Seng, Abdul Hanan Abdullah, Abdul Manan Ahmad	2
162	Technopreneurship as the New Paradigm For E-Business	Prof. Dr. Ahmad Zaki Abu Bakar	5
163	Three-Dimensional Terrain Database Design and Management for Development of Virtual Geographical Information System	Muhamad Najib Zamri, Safie Mat Yatim, Noor Azam Md. Sheriff, Ismail Mat Amin	3
164	Modeling and Simulation of Collision Response Between Deformable Objects	Abdullah Bade, Saandilian Devadas, Daut Daman, Norhaida Mohd Suaib	3
165	Steganography : Hiding Secret Data Into Doubtless Text	Prof Dr. Mohd Aizaini Maarof	2
166	Sound Optimization and Security System using Compression and Encryption Technique	Prof Dr Mohd Aizaini Maarof	2
167	Proxy System Using Squid	Prof Dr. Abd Hanan bin Abdullah	2
168	Feature Selection Method Using Genetic Algorithm for the Classification of Small and High Dimension Data	Mohd Saberi Mohamad, Safaai Deris	1
169	The Crowd Simulation for Interactive Virtual Environments	Muhammad Shafie Abdul Latif, Setyawan Widyarto	4
170	Solving Time Gap Problems through the Optimization of Detecting Stepping	Prof Dr Mohd Aizaini bin Maarof,	2

	Stone Algorithm	Mohd Nizam Omar, Anazida Zainal	
171	Individualizing Learning Material Of Adaptive Hypermedia Learning System Based On Personality Factor (Mbti) Using Fuzzy Logic Techniques	Norreen Binti Haron , Naomie Binti Salim	8
172	Fuzzy Decision Tree for Data Mining of Time Series Stock Market Databases	Mohd Noor Md Sap, Rashid Hafeez Khokhar	1
173	A Multiple Perspectives Review Of Knowledge Management Literature	Dr Rose Alinda Alias	5
174	3D Object Reconstruction and Representation Using Neural Networks	Lim Wen Peng, Siti Mariyam Shamsuddin	1
175	The Development of Feature Extraction And Pattern Matching Techniques For 2D Image For Trademark Logo Recognition	Assoc. Prof. Dr. Dzulkifli bin Mohamad	3
176	A Computerized Handwritten Text Recognition System	Prof. Dr. Ghazali Sulong	3
177	A Computerized Isolated Hand printed Character Recognition System	Prof. Dr. Ghazali bin Sulong	3
178	A Secure Transaction Framework For Client-Server Based E-Commerce	Prof. Dr. Abd. Hanan bin Abdullah	2
179	Malay Spelling Checker & End of Line Word Hyphenation & Database for Encyclopedia Science & Technology Project	Assoc. Prof. Dr. Naomie binti Salim	1
180	Classification and Indexing of 2D Medical Images for Content-Based Retrieval System of Digitized X-Ray Films	Assoc. Prof. Dr. Mohd. Noor bin Md. Sap	1
181	Computerization of Manpower Planning System on Medical Doctor & Specialist in Malaysia	Prof. Dr. Ghazali bin Sulong	3
182	Database Security And Reliability Analysis For Real-time Wireless Update	Assoc. Prof. Dr. Mohd. Noor bin Md. Sap	1
183	Development of a Model for Service Quality information Systems	Prof. Dr. Rose Alinda binti Alias	5
184	Development of Collaborative Environment with Privacy and Conference Control for 3D Protein Structure Visualization	Assoc. Prof. Safie bin Mat Yatim	3
185	Information Systems Planning	Prof. Dr. Rose Alinda binti Alias	7
186	Malaysian IT Technopreneurship Model And Decision Support Tool Kit	Prof. Dr. Ahmad Zaki bin Abu Bakar	5
187	Network Design and Security (NDS)	Prof. Dr. Abd. Hanan bin Abdullah	2
188	Neural-Fuzzy-Ep Application in Scheduling, Planning and Forecasting	Prof. Dr. Safaai bin Deris	1
189	Spatial And Non-Spatial Database Enhancement For Hydrogical Information System (HIS)	Assoc. Prof. Daut bin Daman	3
190	Alternative Negative Selection Framework Of Artificial Immune System For Classification Problems	Associate Professor Dr. Siti Mariyam Shamsuddin	1
191	The Reconstruction Of Sketched Primitive Objects	Dr. Habibollah bin Haron	3
192	An Enhanced Parallel Thinning Algorithm for Handwritten Character Recognition Using Neural Network	Dr Habibollah bin Haron	3
193	Outlier Detection For Breast Cancer Using K-Means And Isodata	PM Dr Mohd Noor Md Sap	1
194	A Analysis of Hierarchical Clustering and Neural Network Clustering in Suggestion Supervisor and Examiner of Thesis	PM Dr Naomie Salim	1
195	An Analysis Of Non Hierarchical And Fuzzy Clustering For Suggestion Of Supervisor And Examiner Of Thesis Title	PM Dr. Naomie Salim	1
196	Developing an Student Performance Evaluation System	Dr. Azizah Abd. Rahman	8
197	Developing A Customer Relationship Management System As A Support Tool To Improve The Services In Perpustakaan Sultanah Zanariah	Dr. Azizah Abdul Rahman	5
198	K-Portal Zakat	Dr Othman Ibrahim	7
199	Redesigning the project monitoring process: Case study at Pejabat Harta Bina UTM	Dr. Azizah Abd. Rahman, Associate Prof. Dr. Rose Alinda Alias	5
200	Comparison Retrieval Schemes Based On Title, Abstract And Bibliography Structures Of Thesis With Different Weighting Schemes	PM Dr Naomie Salim	1
201	Optimization Process Of Numerical Control Code In Manufacturing Endmill Tool Endpoint Sized 20mm	Dr Habibollah bin Haron	6
202	Optimization of Numerical Control Code in Manufacturing Ball End 25mm Tool	Dr Habibollah bin Haron	6
203	Algorithm Enhancement For Host-Based Intrusion Detection System Using Discriminant Analysis	Prof Dr Abdul Hanan bin Abdullah	2
204	Development Of Graphical User Interface (GUI) For Firewall Monitoring System	Prof Dr Abdul Hanan bin Abdullah	2
205	Steganography And Cryptography Apply to Hide X-Ray Image	Prof Dr Aizaini Maarof	2
206	Improved Two-Term Backpropagation Error Function with GA Based Parameter Tuning for Classification Problem	PM Dr. Siti Mariyam Hj Shamsuddin	1

APPENDIX C
Stopword List

a	at	considering
a's	available	contain
able	away	containing
about	awfully	contains
above	b	corresponding
according	be	could
accordingly	became	couldn't
across	because	course
actually	become	currently
after	becomes	d
afterwards	becoming	daily
again	been	date
against	before	definitely
ain't	beforehand	described
all	behind	despite
allow	being	did
allows	believe	didn't
almost	below	different
alone	beside	do
along	besides	does
already	best	doesn't
also	better	doing
although	between	don't
always	beyond	done
am	both	down
among	brief	downwards
amongst	but	during
an	by	e
and	bahru	each
another	c	edu
any	c'mon	e.g.
anybody	c's	eg
anyhow	came	eight
anyone	can	either
anything	can't	else
anyway	cannot	elsewhere
anyways	cant	enough
anywhere	cause	entirely
apart	causes	especially
appear	certain	et
appreciate	certainly	etc
appropriate	changes	even
are	clearly	ever
aren't	co	every
around	com	everybody
as	come	everyone
aside	comes	everything
ask	concerning	everywhere
asking	consequently	ex
associated	consider	exactly

example	hereafter	knows
except	hereby	known
f	herein	last
far	hereupon	lately
few	hers	later
fifth	herself	latter
first	hi	latterly
five	him	least
followed	himself	less
following	his	lest
follows	hither	let
for	hopefully	let's
former	how	like
formerly	howbeit	liked
forth	however	likely
four	i'd	little
from	i'll	look
further	i'm	looking
furthermore	i've	looks
g	ie	ltd
get	if	malaysia
gets	ignored	mainly
getting	immediate	many
given	in	may
gives	inasmuch	maybe
go	inc	me
goes	indeed	mean
going	indicate	meanwhile
gone	indicated	merely
got	indicates	might
gotten	inner	more
greetings	insofar	moreover
h	instead	most
had	into	mostly
hadn't	inward	much
happens	is	must
hardly	isn't	my
has	it	myself
hasn't	it'd	name
have	it'll	namely
haven't	it's	nd
having	its	near
he	itself	nearly
he's	jabatan	necessary
hello	johor	need
help	just	needs
hence	keep	neither
her	keeps	never
here	kept	nevertheless
here's	know	new

next	possible	somebody
nine	presumably	somehow
no	probably	someone
nobody	provides	something
non	que	sometime
none	quite	sometimes
noone	qv	somewhat
nor	rather	somewhere
normally	rd	soon
not	re	sorry
nothing	really	specified
novel	reasonably	specify
now	regarding	specifying
nowadays	regardless	still
nowhere	regards	sub
obviously	relatively	such
of	respectively	sup
off	right	sure
often	said	t's
oh	same	take
ok	saw	taken
okay	say	tell
old	saying	tends
on	says	th
once	second	than
one	secondly	thank
ones	see	thanks
only	seeing	thanx
onto	seem	that
or	seemed	that's
other	seeming	thats
others	seems	the
otherwise	seen	their
ought	self	theirs
our	selves	them
ours	sensible	themselves
ourselves	sent	then
out	serious	thence
outside	seriously	there
over	seven	there's
overall	several	thereafter
own	shall	thereby
particular	she	therefore
particularly	should	therein
pc	shouldn't	theres
per	since	thereupon
perhaps	six	these
placed	smk	they
please	so	they'd
plus	some	they'll

they're	vs	would
they've	want	wouldn't
think	wants	yes
third	was	yet
this	wasn't	you
thorough	way	you'd
thoroughly	we	you'll
those	we'd	you're
though	we'll	you've
three	we're	your
through	we've	yours
throughout	welcome	yourself
thru	well	yourselves
thus	went	zero
to	were	
today	weren't	
together	what	
too	what's	
took	whatever	
toward	when	
towards	whence	
tried	whenever	
tries	where	
truly	where's	
try	whereafter	
trying	whereas	
twice	whereby	
two	wherein	
un	whereupon	
under	wherever	
unfortunately	whether	
unless	which	
unlikely	while	
until	whither	
unto	who	
up	who's	
upon	whoever	
us	whole	
use	whom	
used	whose	
useful	why	
uses	will	
using	willing	
usually	wish	
uucp	with	
value	within	
various	without	
very	won't	
via	wonder	
viz	would	

APPENDIX D
Porter Stemming Rule

Step 1a Rules

Conditions	Suffix	Replacement	Examples
NULL	sses	ss	caresses -> caress
NULL	ies	i	ponies -> poni ties -> tie
NULL	ss	ss	carress -> carress
NULL	s	NULL	cats -> cat

Step 1b Rules

Conditions	Suffix	Replacement	Examples
(m>0)	ced	ce	feed -> feed agreed -> agree
(*v*)	ed	NULL	plastered -> plaster bled -> bled
(*v*)	ing	NULL	motoring -> motor sing -> sing

Step 2 Rules

Conditions	Suffix	Replacement	Examples
(m>0)	ational	ate	relational -> relate
(m>0)	tional	tion	conditional -> condition rational -> rational
(m>0)	enci	ence	valenci -> valence
(m>0)	anci	ance	hesitanci -> hesitance
(m>0)	izer	ize	digitizer -> digitize
(m>0)	abli	able	conformabli -> conformable
(m>0)	alli	al	radicalli -> radical
(m>0)	entli	ent	differentli -> different
(m>0)	eli	e	vileli -> vile
(m>0)	ousli	ous	analogousli -> analogous
(m>0)	ization	ize	vietnamization -> vietnamize
(m>0)	ation	ate	predication -> predicate
(m>0)	ator	ate	operator -> operate
(m>0)	atism	al	feudalism -> feudal
(m>0)	iveness	ive	decisiveness -> decisive
(m>0)	fulness	ful	hopefulness -> hopeful
(m>0)	ousness	ous	callousness -> callous
(m>0)	aliti	al	formaliti -> formal
(m>0)	iviti	ive	sensitiviti -> sensitive
(m>0)	biliti	ble	sensibiliti -> sensible

Step 3 Rules

Conditions	Suffix	Replacement	Examples
(m>0)	icate	ic	triplicate -> triplic
(m>0)	ative	NULL	formative -> form
(m>0)	alize	al	formalize -> formal
(m>0)	iciti	ic	electriciti -> electric
(m>0)	ical	ic	electrical -> electric
(m>0)	ful	NULL	hopeful -> hope
(m>0)	ness	NULL	goodness -> good

Step 4 Rules

Conditions	Suffix	Replacement	Examples
(m>1)	al	NULL	revival -> reviv
(m>1)	ance	NULL	allowance -> allow
(m>1)	ence	NULL	inference -> infer
(m>1)	er	NULL	airliner -> airlin
(m>1)	ic	NULL	gyroscopic -> gyroscop
(m>1)	able	NULL	adjustable -> adjust
(m>1)	ible	NULL	defensible -> defens
(m>1)	ant	NULL	irritant -> irrit
(m>1)	ement	NULL	replacement -> replac
(m>1)	ment	NULL	adjustment -> adjust
(m>1)	ent	NULL	dependent -> depend
(m>1 and (*<S> or *<T>))	ion	NULL	adoption->adopt
(m>1)	ou	NULL	homologou->homolog
(m>1)	ism	NULL	communism->commun
(m>1)	ate	NULL	activate->activ
(m>1)	iti	NULL	angulariti->angular
(m>1)	ous	NULL	homologous ->homolog
(m>1)	ive	NULL	effective->effect
(m>1)	ize	NULL	bowdlerize->bowdler

Step 5a Rules

Conditions	Suffix	Replacement	Examples
(m>1)	e	NULL	probate -> probat rate -> rate
(m=1 and not *o)	e	NULL	cease -> ceas

Step 5b Rules

Conditions	Suffix	Replacement	Examples
(m>1 and *d and *<L>)	NULL	single letter	control -> control roll -> roll

APPENDIX E
Preprocessing Result

No	Title	Supervisor	ExpertSurvey
1	COMPARATIVE#1 STUDY#1 SMART#1 SCHOOL#2 NORMAL#1 USAGE#1 ICT#1	Prof. Dr. Rose Alinda Binti Alias	5
2	DECISION#1 SUPPORT#1 SYSTEM#1 BASED#1 NEURAL#1 NETWORKS#1 SELECT#1 CANDIDATES#1 COMPUTER#1 SCIENCE#1 DEGREE#1 PROGRAM#1	Prof.Madya Dr.Mohd Noor Bin Md Sap	1
3	Prototyp#1 Virtual#1 Intellig#1 Map#1 System#1	Prof Dr Ahmad Zaki b Abu Bakar	5
4	Resourc#1 usag#1 analyz#1 distribut#1 threat#1 simul#1 intrus#1 detect#1 system#1 hose#1	PM Dr Mohd Azaini b Maarof	2
5	Active#1 Reaction#1 analyz#1 distribut#1 threat#1 simul#1 parallel#1 Intrusion#1 Respons#1 system#1	PM Dr Mohd. Aizaini Maarof	2
6	Knowledg#1 Manag#1 Applicat#1 Support#1 Commun#1 Practic#1 High#1 Level#1 Learn#1 Institut#1	Prof. Madya Dr. Shamsul bin Shahibuddin, En. Mohammad Nazir bin Ahmad @ Sharif	5
7	Applicat#1 Fuzzi#1 Logic#1 Decis#1 Support#1 System#1 Select#1 Subject#1 Univers#1	PM. Abd. Manan Ahmad	1
8	CLASS#1 TIMETABLING#1 MODIFIED#1 GENETIC#1 ALGORITHM#1	PROF. MADYA DR. SAFAAI BIN DERIS	1
9	Comput#1 Control#1 System#1	P.M. Dr. Mohd Aizaini Bin Maarof	2
10	Content#1 Manag#1 Framework#1 Malaysian#1 Govern#1 Web#1 Site#1	PROF. ZAMRI B. MOHAMED	5
11	Critic#1 Success#1 Factor#1 Manag#1 Dot.Com#1 Compani#1	PM DR HARIHODIN B SELAMAT	5
12	Denial#1 Servic#1 Attack#1 Detect#1	Prof Dr Abdul Hanan b Abdullah	2
13	Develop#1 Univers#1 Knowledg#1 Commun#1 Sm#1 Technolog#1	PM Dr. Rose Alinda Alias, Professor Dr. Ahmad Zaki bin Abu Bakar	5
14	Develop#1 Web#1 Base#1 Tourism#2 Informat#1 System#1 Sarawak#1 Board#1	Prof Zamri b Mohamed	5
15	Develop#1 Prototyp#1 Johor#1 Tourism#1 Informat#1 System#1 Web#1 Environ#1	PM Dr. Mohd. Noor bin Md. Sap	5
16	Easi#1 Link#1 Informat#1 Centr#1 Administr#1 (Elica)#1	PM Dr Shamsul b Sahibuddin	5
17	Electron#1 Commerc#1 Comput#1 Shop#1	PM Dr Harihodin b Selamat	7
18	Enhanc#1 Custom#1 Relationship#1 Process#1 Southern#1 Sumatra#1 District#1 Office#1 P.T#1 Telekomunikasi#1 Indonesia#1 Tbk#1 Integrat#1 Multipl#1 Commun#1 Servic#1	PM Dr Abdul Samad b Hj Ismai	4
19	Informat#1 Commun#1 Technolog#1 Manag#1 Guidelin#1 Small#1 medium#1 Industri#1 rural#1 area#1	PROFESSOR ZAMRI BIN MOHAMED, Encik Md. Hafiz Bin Selamat, Encik Mohd. Zaidi Bin Abdul Rozan	5
20	Beam#1 Search#1 Implement#1 Solv#1 Person#1 Comput#1 Configur#1 Problem#1	PM Abdul Manan Ahmad	1
21	IMPLEMENTATION#1 CONSTRAINT#1 BASED#1 SCHEDULING#1 NURSE#1 SHIFT#1 CRYSTAL#1 WARD#1 HOSPITAL#1 UNIVERSITY#1 (HUSM)"#1	Professor Dr. Safaai Bin Deris, PM Safie Bin Mat Yatim, PM Abdul Manan	1
22	Informat#1 Secur#1 Polici#1 Univers#1 Teknolog#1 MARA#1	Prof Dr Abdul Hanan b Abdullah	2
23	Integrat#1 Workspac#1 Aware#1 Collabor#1 Case#1 Tool#1	PM DR. Shamsul Sahibuddin, Pn Mazleena	4
24	Islamic#1 E-Organiz#1	PM Dr Shamsul b Sahibuddin	5
25	Guidelin#1 Form#1 Improve#1 Effectiv#1 Supervis#1 control#1 ICT#1 Project#2 govern#1 sector#1 Leader#1 Committe#1	Prof Zamri b Mohamed	7
26	Reengin#1 Magazin#1 Ordere#1 System#2 Replac#1 Structur#1 Analysi#1 Object#1 Orient#1 Design#1 UML#1	PM Dr Mohd Noor b Md Sap	1
27	Knowledg#2 Cent#1 Contribut#1 Metric#1 Manag#1	Prof Dr Ahmad Zaki b Abu Bakar, En Mohd Zaidi b Abdul Rozan	5
28	Manag#1 Tutor#1 System#1 Propag#1 Plants-Liv#1 Skill#1 Subject#1	PM Dr Rahman b Ahmad	8
29	Mobil#1 Protect#1 System#1	Assoc. Prof. Dr. Shamsul Bin Sahibuddin	4
30	Neural#1 Fuzzi#1 Handwritten#1 Digit#1 Recognit#1	PM Dr Siti Mariyam bt Hj Shamsudin	3
31	Develop#1 Comput#1 Game#1 Engine#1 simul#1 strategi#1 style#1	Prof. Dr. Ahmad Zaki Abu Bakar	5
32	FSKSM#1 Sharenet#1 Implement#1 creat#1 knowledg#1 share#1 cultur#1	PM Abdul Manan b Ahmad	1
33	Prototyp#1 Develop#1 Plan#1 Informat#1 System#1 Techniqu#1 Tool#1	PM Dr Rose Alinda Alias	7
34	Develop#1 Strateg#1 Manag#1 Assessment#1 System#1 IPTS#1 weightag#1 method#1 :#1 ITP-YPJ#1	PM DR ROSE ALINDA BT Alias, Pn Zeti Darleena	5
35	Develop#1 Object#1 Orient#1 IKS#1 Administr#1 System#1	PM Dr Rose Alinda bt Alias	5
36	Implement#1 Infomediari#1 Concept#1 E-Learn#1 ENviron#1	PM Dr Rahman b Ahmad	8
37	Hypermedia#1 Applicat#1 Model#1 Combin#1 UML#1 HDM#1	Prof Dr Safaai b Deris	1
38	Recurr#1 Neural#1 Network#1 Predict#1 Hous#1 Price#1	PM Dr Siti Mariyam bt Hj Shamsudin	1
39	Enhancement#1 MOO#1 Tool#1 Distanc#1 Learn#1	PM Dr. Shamsul Sahibuddin	7
40	Effectiv#1 Hybrid#1 Bound#1 Volum#1 Seri#1 Strategi#1 Detect#1 Collis#1 N#1 Rigid#1 Convex#1 Object#1 Interact#1 Virtual#1 Environmen#1	PM Daut Daman	3
41	Speech#1 Recognit#1 Isolat#1 Word#1 Sampl#1 Support#1 Vector#1 Machin#1	PM DR SITI MARIAM BINTI HJ	3

		SHAMSUDIN, EN MD SAH BIN HJ SALAM	
42	Grai#1 Scale#1 Fingerprint#1 Image#1 Minutia#1 Detect#1 Ridg#1 Line#1 Algorithm#1	Prof Dr Ghazali Sulong	3
43	Neural#1 Network#1 time#1 seri#1 forecast#1	PM Dr Salihin b Ngadiman	1
44	Implement#1 Knowledg#1 Manag#1 Higher#1 Learn#1 Institut#1 Case#1 studi#1 FSKSM#1	PM Dr Shamsul b Sahibuddin, Pn Norhawaniah bt Zakaria	5
45	Rainfal#1 Forecast#1 Neural#1 Network#1 base#1 Massiv#1 Meteorolog#1 Data#1	P.M. Dr. Mohd Noor bin Md. Sap	1
46	Comparison#1 Classif#1 Result#1 Undiscret#1 Discret#1 Data#1 Back#1 Propag#1 Rough#1 Set#1	PM Dr Siti Mariyam bt Hj Shamsudin	1
47	Neural#1 Network#1 Optimiz#1 Genet#1 Algorithm#1 Speech#1 recognit#1	PM Dr Zulkifli Mohamad	3
48	Enhancement#1 Dick#1 Carei#2 (2001)#1 Model#2 Step#1 develop#1 intellig#1 learn#1 system#1	PM Noraniah bt Mohd Yassin	8
49	Informat#1 system#1 Qualiti#1 Servic#1 Assessment#1 :#1 Custom#1 Autom#1 Support#1 System#1 (CASS)#1 Telekom#1	PROF MADYA DR ROSE ALINDA BTE ALIAS	5
50	Comparison#1 linear#1 summat#1 techniqu#1 neural#1 network#1 model#1 decis#1 support#1 student#1 intak#1 Higher#1 Educat#1 Institut#1	PM Dr Mohd Noor b Md Sap	1
51	Perform#1 Comparison#1 Java#1 RMI#1 CORBA#1 Multi-Level#1 Market#1 Busi#1 Structur#1	PM. DR. Rose Alinda Alias	5
52	Collabor#1 environ#1 JPA#1 top#1 manag#1 :#1 Manag#1 Profession#1	PM Dr Shamsul b Sahibuddin	5
53	Properti#1 Consult#1 Computer#1	Pm Abdul Manan b Ahmad	1
54	Knowledg#1 Audit#1 Portal#1 Public#1 Higher#1 Educat#1	PM Dr Rose Alinda Alias	5
55	Small#1 Medium#1 Industri#1 Commun#1 Portal#1	PM Safie Mat Yatim, En Norhashim Abu Samah	5
56	UPSI#1 Educat#1 Commun#1 Portal#1	PM Dr Shamsul Sahibuddin	5
57	Knowledg#1 Manag#1 System#1	Prof Dr Ahmad Zaki b Abu Bakar	5
58	Seminar#1 Manag#1 Monitor#1 Portal#1	Prof Dr Safaai b Deris	1
59	PRODUCT#1 INFORMATION#1 SEARCHING#1 WAP#1	PM Dr Mohd Azaini b Maarof	2
60	Project#1 Time#1 Manag#1 Commun#1 System#1	Prof Dr Safaai b Deris	1
61	Protein#1 Sekondari#1 Structur#1 Predict#1 Amino#1 Acid#1 Sequenc#1 Neural#1 Network#1 Classifi#1 Base#1 Dempster#1 Shafer#1 Theori#1	Prof Dr Safaai b Deris, PM Dr Rosli Md Illias	1
62	Evaluat#1 System#1 E-Learn#1 Portal#1	PM Dr Mohd Noor b Md Sap	5
63	World#1 Islamic#1 Trade#1 Busi#1 Game#1 (WITNES)#1 Prototyp#1	En Noor Azam b Mohd Sheriff	3
64	Prototyp#1 System#1 Person#1 Firewall#1	Prof. Dr. Abdul Hanan Bin Abdullah	2
65	Prototyp#1 Profit#1 Loss#1 Statement#1 Analysi#1 Simul#1 Model#1 Fuzzi#1 Logic#1	Prof Dr Ahmad Zaki b Abu Bakar	5
66	Region#1 Base#1 Digit#1 Image#1 Segment#1 Pixel#2 Caste-Mark#1 Discrimin#1	PM Dr Mohd Noor b Md Sap, Pm Dr Harihodin b Selamat	1
67	Tool#1 Design#1 Web#1 Base#1 Support#1 Collabor#1 Learn#1	PM Dr Harihodin b Selamat	5
68	Trap#1 detach#1 design#1 Simpl#1 Network#1 Manag#1 Protocol#1 (SNMP)#1 wireless#1 applic#1 protocol#1 (WAP)#1 Client#1 Server#1 System#1	Prof Dr. Shamsul Sahibuddin	4
69	Intellig#1 Mobil#1 Agent#1 Open#1 Architectur#1 Distribut#1 Applicat#1	Prof Dr Safaai b Deris	1
70	Decis#1 Support#1 System#1 Neural#1 Network#1 Bank#1 Loan#1 Applicat#1	PM Dr Mohd Noor Md Sap	1
71	School#1 Disciplin#1 Decis#1 Support#1 System#1	PM Abdul Manan b Ahmad	1
72	decis#1 support#1 system#2 servic#1 qualiti#1 inform#1 IPTS#1	PM Dr Rose Alinda bt Alias	5
73	Agenci#1 Manag#1 Smart#1 System#1	PM. SAFIE MAT YATIM, PM. DR. SAFAAI DERIS	5
74	Qualiti#1 Assuranc#1 System#1 Top#1 Empire#1 Sdn#1 Bhd#1	Prof. Madya Dr. Ab. Rahman Ahmad	6
75	Softwar#1 Standard#1 Control#1 System#1	Prof Dr. Hanan b Abdullah	2
76	Assembl#1 Line#1 Balanc#1 Workstat#1 System#1 Heurist#1 Method#1	Prof. Madya Dr. Ab Rahman Ahmad, Dr. Masine bte Md Tap	3
77	Student#1 Informat#1 System#1	En Nadzari b Shaari, Prof Dr Safaai b Deris	5
78	student#1 disciplin#1 system#1 school#1 involv#1 merit#1 demerit#1 process#1	PM Abdul Manan b Ahmad	1
79	Electron#1 Document#1 Deliveri#1 System#1	PM Dr Mohd Noor b Md Sap	5
80	Data#1 Recoveri#1 System#2 Disc#1 Forens#1 Window#1 Operat#1	PROF MADYA DR. SHAMSUL SAHIBUDDIN	5
81	Intellig#1 tutor#1 system#1 :#1 Possibl#1 Statist#1 Topic#1	PM Noraniah bt Mohd Yassin	8
82	Faraid#1 knowledg#1 inform#1 technolog#1 Base#1 Web#1	PM Abdul Manan Ahmad	5
83	Staff#1 Informat#1 Manag#1 System#1 Base#1 Lotu#1 Note#1 :#1 Mydin#1 Mohammad#1 Son#1 Sdn#1 Bhd#1	Dr Shamsul b Sahibuddin	5
84	Knowledg#1 Manag#1 System#1 Solv#1 ISO#1 9000#1	Dr. Azizah Binti Abdul Rahman, En. Azlan Bin Mohd Zain	5
85	Student#1 Perform#1 Assessment#1 System#1	PM Noraniah Mohd.Yassin	8

86	Online#1 Vehicl#1 Sale#1 System#1	P.M. Safie Mat Yatim, P.M. Dr. Mohd. Noor Md. Sap	5
87	develop#1 incom#2 statement#1 individu#1 tax#1 calcul#1 on-lin#1 system#1	PM Dr Shamsul b Sahibuddin	5
88	Custom#1 Profil#1 System#1 TELEKOM#1 BERHAD#1	Prof. Madya Dr. Harihodin Selamat	7
89	Decis#1 Support#1 System#1 Vehicl#1 Bui#1 Plan#1 Hire#1 Purchas#1	PM Dr Shamsul b Sahibuddin, Pn Norhawaniah bt Hj Zakaria	5
90	APPLICATION#1 ENHANCED#1 GENETIC#1 ALGORITHM#1 CLASS#1 TIMETABLING#1 PROBLEM#1	Prof. Dr. Safaai bin Deris	1
91	Effect#1 Malaysian#1 Smart#1 School#1 Public#1 Univer#1 Curriculum#1 Structur#1 Term#1 Basic#1 Subject#1	PM Abdul Manan b Ahmad	1
92	RETRACE#1 TRAVELING#2 SALESMAN#2 PROBLEM:#1 EXTENSION#1 PROBLEM#1	PROF. MADYA DR. SAFAAI BIN DERIS	1
93	Trademark#1 Match#1 Algorithm#1 base#1 Simplifi#1 Featur#1 Extraction#1	Dr Dzulkifli Mohamad	3
94	Visual#1 Larg#1 Data#1 Set#1 Triangular#1 DTM#1	PM Daut Daman	3
95	Busi#1 Advertis#1 Web#1	PM Dr Mohd Azaini b Maarof	2
96	Workflow#1 Manag#1 System#1 Strata#1 Titl#1 Applicat#1 Feder#1 Land#1 Mine#1 Depart#1	PM Dr Rose Alinda bt Alias	5
97	System#1 Analysi#1 Comparison#1 Data#1 Flow#1 Diagram#1 Case#1	PM Dr Rose Alinda Binti Alias	5
98	Recognit#1 Plate#1 number#1 locat#1 statist#1 method#1	Dr. Dzulkifli Mohamad	3
99	Transit#1 System#1 Design#1 Function#1 Orient#2 Object#1	Pn Azizah binti Abdul Rahman, Pn Nor Hawaniah binti Zakaria	5
100	Effect#1 Architectur#1 Exact#1 Forecast#1 Backpropag#1	Prof Dr. Safaai bin Deris	1
101	Electron#1 Commerci#1 System#1 base#1 Letter#1 Credit#1	PM. Dr. Abdul Hannan Abdullah	2
102	Build#1 Prototyp#1 Data#1 Warehous#1 Case#1 Studi#1 FAMA#1	PM Dr. Abdul Hannan Abdullah	2
103	Leadership#1 Knowledg#1 Manag#1 Portal:#1 Prototyp#1	Prof. Dr. Ahmad Zaki bin Abu Bakar	5
104	Explore#1 Notion#1 Servic#3 Assessment#1 Context#1 Informat#1 System#1 Quality#1 (ISSQ)#1 Malaysian#1 Public#1	Associate Professor Dr. Rose Alinda binti Alias	5
105	Decis#1 Support#1 System#1 Rural#1 Digit#1 Divid#1 Program#1	Prof. Zamri bin Mohamed	7
106	Develop#1 Web#1 Creation#1 Manag#1 Tool#1	PM Noraniah Mohd Yassin	8
107	Develop#1 MSCIT#1 Learn#1 Portal#1 Prototyp#1 Case#1 Studi#1 FSKSM#1 MSc#1 Programm#1	Dr. Azizah binti Abdul Rahman	5
108	Prototyp#1 Flood#2 Manag#1 Support#1 System#1 Nation#1 Forecast#1 Center#1 Depart#1 Irrigat#1 Drainag#1	PM Dr. Rose Alinda binti Alias	5
109	Enhanc#1 Decis#1 Make#1 Process#1 Project#1 Monitor#1 Environ#1 Plan#1 Develop#1 Divis#1 Ministri#1 Health#1	Professor Zamri bin Mohamed	7
110	Manag#1 Informat#1 System#1 Equipment#1 Efficienc#1 OEE#1 Analysi#1 Decis#1 Make#1	PM Dr. Safaai bin Deris	1
111	Health#1 Promot#1 Portal#1	PM Dr. Shamsul bin Shahibuddin	5
112	Informat#1 System#1 Plan#1 Secondari#1 School#1 :#1 Case#1 Studi#1 SMK#1 Mutiara#1 Rini#1 Skudai#1	PM Dr. Rose Alinda binti Alias	7
113	Person#1 Budget#1 Decis#1 Support#1 System#1	PM Dr. Shamsul bin Sahibuddin	5
114	Electron#1 Claim#1 Manag#1 System#1	PM Abdul Manan Ahmad	1
115	Comput#1 Assist#1 Learn#1 -#1 Algebra#1 Fraction#1 EIF#1 AProach#1	PM Dr. Ab. Rahman bin Ahmad	8
116	Pornograph#1 web#1 page#1 filter#1 system#1 Neural#1 Network#1 Model#1	PM Dr. Siti Mariyam Hj. Shamsuddin	1
117	Zakat#1 Payment#1 System#1 Jabatan#1 Agama#1 Islam#1	PM Dr. Safaai bin Deris	1
118	Mobil#1 Wireless#1 Ward#1 Hand#1 System#1 WiH#1	PM Dr Rose Alinda bin Alias	5
119	Integrat#1 School#1 Manag#1 Informat#1 System#1	PM Safie bin Mat Yatim	8
120	Smart#1 Kindergarten#1 Manag#1 System#1	Dr. Muhammad Shafie Hj Abdul Latiff	4
121	Institut#1 Knowledg#1 Share#1 Senior#1 Officer#1 Prison#1 Depart#1	Professor Zamri bin Mohamed	7
122	Time#1 Tabl#1 Schedul#1 System#1 Primari#1 School#1	PM Abdul Manan Ahmad	1
123	Activiti#1 Base#1 Cost#1 Softwar#1 Manufactur#1 Industri#1	Dr. Muhammad Shafie Abdul Latiff	4
124	Knowledg#1 Classif#1 System#1 Sistem#1 Saraan#1	Prof. Dr. Ahmad Zaki bin Abu Bakar	5
125	Online#1 Data#1 Stora#1 :#1 Drivepod#1	PM Dr. Rose Alindabinti Alias	5
126	Busi#1 Develop#1 Autom#1 Market#1 Basket#1 Analysi#1	PM Abdul Manan Ahmad	1
127	Academ#1 Advisor#1 Expert#1 System#1	PM Dr. Safaai bin Deris	1
128	Teacher#1 Perfom#1 Apraisal#1 System#1	PM Dr. Mohd Noor bin Md. Sap	1
129	Acquisit#1 Online#1 System#1 PSZ#1	PM Dr. Mohd Noor bin Md. Sap	1
130	Design#1 implement#1 Doubl#1 Cube#1 Data#1 Model#1	PM Dr. Harihodin bin Selamat, PM Daut Daman	7
131	Text#1 Block#1 Index#1 Informat#1 Retriev#1	PM Safie bin Mat Yatim, PM Sarudin bin Bakri	3

132	Informat#1 Secur#1 Polici#1 Develop#1 Jabatan#1 Kastam#1 Diraja#1	PM Dr. Mohd. Aizaini Maarof	2
133	Cluster-Bas#1 Compuond#1 Select#1 Fuzzi#1 Cluster#1	PM Dr. Naomie binti Salim, Dr Ali bin Selamat	1
134	Online#1 Journal#2 Manag#1 System#1 Informat#1 Technolog#1	PM Dr. Naomie bin Salim	1
135	Analysi#1 Cluster#1 Chemic#1 Data#1 Genet#1 Algorithm#1	PM Dr. Naomie binti Salim, Dr. Ali bin Selamat	1
136	Task#1 Monitor#1 Product#1 Manag#1 System#1 (Case#1 Study:#1 SPMB#1 Workshop)#1	Dr. Azizah binti Abdul Rahman	5
137	Decis#1 Support#1 System#1 Assign#1 Machine#1 Top#1 Empire#1 Industri#1 Sdn#1 Bhd#1	PM Dr. Abd. Rahman bin Ahmad	6
138	Web#1 Base#1 Job#1 Applicat#1 System#1	PM Dr. Shamsul Sahibuddin	5
139	Predict#1 Life#1 Expectanc#1 Patient#1 Hepat#1 Support#1 Vector#1 Machin#1 Wrapper#1 Method#1	Professor Dr. Safaai bin Deris	1
140	Rectifi#1 Lot#1 Size#1 Method#1 Forward#1 Wagner#1 Whitin#1 Roll#1 Horizon#1 Environ#1	PM Dr. Mohd Salihin bin Ngadiman	1
141	Assessment#1 Perform#1 Candid#1 Find#1 System#1	Prof. Dr. Ahmad Zaki bin Abu Bakar	8
142	Genet#1 Algorithm#1 Direct#1 Mutat#1 Solv#1 Timet#1 Problem#1	Assoc. Prof. Dr. Mohd Salihin bin Ngadiman, Puan Roselina binti Sallehudin	1
143	Human#1 Animat#1 Neural#1 Network#1	PM Dr. Siti Mariyam binti Shamsudin	1
144	Comparison#1 Effectiv#1 Probabl#1 Model#2 Vector#1 Space#1 Compound#1 Similar#1 Search#1	PM Dr. Naomie binti Salim, Puan Razana Alwee	1
145	Measur#1 System#1 Analysi#1 MSA#1 Automot#1 Manufactur#1 Industri#1 GR#1 R#1	PM Dr. Mohd Salihin bin Ngadiman	6
146	Bioactiv#1 Classif#1 Anti#1 AIDS#1 Compound#1 Neural#1 Network#1 Support#1 Vector#1 Machine:#1 Comparison#1	Assoc. Prof. Dr. Naomie binti Salim	1
147	Identifi#1 Molecul#1 Bioactiv#1 AIDS#1 :#1 Comparison#1 Rough#1 Set#1 Neural#1 Network#1	PM Dr. Naomie binti Salim	1
148	Find#1 Coeffici#2 Fusion#1 Similar#1 Search#1 Neural#1 Network#1 Algorithm#1	PM Dr. Naomie binti Salim	1
149	Pairwis#1 sequenc#2 align#2 select#1 effect#1 substitut#1 matric#1 gap#1 penalti#1 paramet#1 dynam#1 Program#1	PM Dr. Naomie binti Salim, Encik Muhamad Razib bin Othman	1
150	Promot#1 Reflect#1 Practic#1 UTM#1 Teach#1 Commun#1 User#1 Informat#1 Technolog#1	PM Dr. Rose Alinda binti Alias, PM Dr. Abdul Samad bin Ismail	5
151	Prototyp#1 Learn#1 assess#1 applic#1 base#1 Bloom#1 Taxonomi#1 Physic#1 Form#1 4#1	PM Dr. Mohd Noor bin Md. Sap	8
152	Knowledg#1 Manag#2 System#1 Rosettanet#1 Implement#1 Johor#1	Prof. Dr. Ahmad Zaki bin Abu Bakar, En. Md. Hafiz bin Selamat	5
153	Rain#1 Distribut#1 Cluster#1 Data#1 Mine#1 :#1 Comparison#1 Associat#1 Rule#1 Techniqu#1 Statist#1 Method#1	PM Dr. Mohd. Noor bin Md. Sap	1
154	Studi#1 Entrepreneur#1 Intention#1 Informat#1 technolog#1 Technopreneur#1	Prof. Dr. Ahmad Zaki bin Abu Bakar	5
155	Featur#1 Extraction#1 Protein#1 Homolog#1 Detect#1 Hidden#1 Markov#1 Model#1 Combin#1 Score#1	Nazar M. Zaki, Safaai Deris , Rosli M. Illias	1
156	Electric#1 Applianc#1 Control#1 System#1 Internet#1 Base#1 Parallel#1 Port#1	Prof Dr. Abdul Hanan bin Abdullah	2
157	Develop#1 Surfac#1 Reconstruct#1 Ship#1 Hull#1 Design#1	Fadni Bin Forkan, Mahmoud Ali Ahmed, Ang Swee Wen, Siti Mariyam Hj. Shamsuddin, Cik Suhaimi Bin Yusof, Mohd. Razak Samingan, Yahya Samian	3
158	CSCW#1 System#1 Office#1 Environ#1 Applicat#1	Prof Dr Mohd Aizaini Maarof	2
159	Fingerprint#1 Classif#1 Approaches:#1 Overview#1	Leong Chung Ern, Dr. Ghazali Sulong	3
160	Hybrid#1 Trust#1 Manag#1 Model#1 MAS#1 Base#1 Trade#1 Societi#1	Prof Dr. Aizaini Maarof, Krishna K.2	2
161	Interact#1 Agent#1 (Argu#1 Cooper#1 Agents)#1	Ng Kee Seng, Abdul Hanan Abdullah, Abdul Manan Ahmad	2
162	Technopreneurship#1 Paradigm#1 E-Busi#1	Prof. Dr. Ahmad Zaki Abu Bakar	5
163	Dimension#1 Terrain#1 Databas#1 Design#1 Manag#1 Develop#1 Virtual#1 Geograph#1 Informat#1 System#1	Muhamad Najib Zamri, Safie Mat Yatim, Noor Azam Md. Sheriff, Ismail Mat Amin	3
164	Model#1 Simul#1 Collis#1 Respons#1 Deform#1 Object#1	Abdullah Bade, Saandilian Devadas, Daut Daman, Norhaida Mohd Suaib	3
165	Steganographi#1 :#1 Hide#1 Secret#1 Data#1 Doubtless#1 Text#1	Prof Dr. Mohd Aizaini Maarof	2
166	Sound#1 Optimiz#1 Secur#1 System#1 Compress#1 Encryption#1 Techniqu#1	Prof Dr Mohd Aizaini Maarof	2
167	Proxi#1 System#1 Squid#1	Prof Dr. Abd Hanan bin Abdullah	2
168	Featur#1 Select#1 Method#1 Genet#1 Algorithm#1 Classif#1 Small#1 High#1 Dimens#1 Data#1	Mohd Saberi Mohamad, Safaai Deris	1
169	Crowd#1 Simul#1 Interact#1 Virtual#1 Environ#1	Muhammad Shafie Abdul Latif, Setyawan Widyarto	4
170	Solv#1 Time#1 Gap#1 Problem#1 Optimiz#1 Detect#1 Step#1 Stone#1 Algorithm#1	Prof Dr Mohd Aizaini bin Maarof, Mohd Nizam Omar, Anazida Zainal	2
171	Individu#1 Learn#2 Materi#1 Adaptiv#1 Hypermedia#1 System#1 Base#1 Person#1 Factor#1 Mbti#1 Fuzzi#1 Logic#1 Techniqu#1	Norreen Binti Haron , Naomie Binti Salim	8

172	Fuzzi#1 Decis#1 Tree#1 Data#1 Mine#1 Time#1 Seri#1 Stock#1 Market#1 Databas#1	Mohd Noor Md Sap, Rashid Hafeez Khokhar	1
173	Multipl#1 Perspect#1 Review#1 Knowledg#1 Manag#1 Literatur#1	Dr Rose Alinda Alias	5
174	3D#1 Object#1 Reconstruct#1 Represent#1 Neural#1 Network#1	Lim Wen Peng, Siti Mariyam Shamsuddin	1
175	Develop#1 Featur#1 Extraction#1 Pattern#1 Match#1 Techniqu#1 2D#1 Image#1 Trademark#1 Logo#1 Recognit#1	Assoc. Prof. Dr. Dzulkifli bin Mohamad	3
176	Computer#1 Handwritten#1 Text#1 Recognit#1 System#1	Prof. Dr. Ghazali Sulong	3
177	Computer#1 Isolati#1 Hand#1 print#1 Charact#1 Recognit#1 System#1	Prof. Dr. Ghazali bin Sulong	3
178	Secur#1 Transact#1 Framework#1 Client-Serv#1 Base#1 E-Commerc#1	Prof. Dr. Abd. Hanan bin Abdullah	2
179	Malai#1 Spell#1 Checker#1 End#1 Line#1 Word#1 Hyphen#1 Databas#1 Encyclopedia#1 Scienc#1 Technolog#1 Project#1	Assoc. Prof. Dr. Naomie binti Salim	1
180	Classif#1 Index#1 2D#1 Medic#1 Image#1 Content#1 Base#1 Retriev#1 System#1 Digit#1 X#1 Rai#1 Film#1	Assoc. Prof. Dr. Mohd. Noor bin Md. Sap	1
181	Computer#1 Manpow#1 Plan#1 System#1 Medic#1 Doctor#1 Specialist#1	Prof. Dr. Ghazali bin Sulong	3
182	Databas#1 Secur#1 Reliabl#1 Analysi#1 Real-tim#1 Wireless#1 Update#1	Assoc. Prof. Dr. Mohd. Noor bin Md. Sap	1
183	Develop#1 Model#1 Servic#1 Qualiti#1 inform#1 System#1	Prof. Dr. Rose Alinda binti Alias	5
184	Develop#1 Collabor#1 Environ#1 Privaci#1 Confer#1 Control#1 3D#1 Protein#1 Structur#1 Visual#1	Assoc. Prof. Safie bin Mat Yatim	3
185	Informat#1 System#1 Plan#1	Prof. Dr. Rose Alinda binti Alias	7
186	Malaysian#1 Technopreneurship#1 Model#1 Decis#1 Support#1 Tool#1 Kit#1	Prof. Dr. Ahmad Zaki bin Abu Bakar	5
187	Network#1 Design#1 Secur#1 (NDS)#1	Prof. Dr. Abd. Hanan bin Abdullah	2
188	Neural#1 Fuzzi#1 Ep#1 Applicat#1 Schedul#1 Plan#1 Forecast#1	Prof. Dr. Safaai bin Deris	1
189	Spatial#1 Non-Spati#1 Databas#1 Enhancement#1 Hydrogil#1 Informat#1 System#1 (HIS)#1	Assoc. Prof. Daut bin Daman	3
190	Altern#1 Neg#1 Select#1 Framework#1 Artifici#1 Immune#1 System#1 Classif#1 Problem#1	Associate Professor Dr. Siti Mariyam Shamsuddin	1
191	Reconstruct#1 Sketch#1 Primit#1 Object#1	Dr. Habibollah bin Haron	3
192	Enhanc#1 Parallel#1 Thin#1 Algorithm#1 Handwritten#1 Charact#1 Recognit#1 Neural#1 Network#1	Dr Habibollah bin Haron	3
193	Outlier#1 Detect#1 Breast#1 Cancer#1 K-Mean#1 Isodata#1	PM Dr Mohd Noor Md Sap	1
194	Analysi#1 Hierarch#1 Cluster#2 Neural#1 Network#1 Suggest#1 Supervisor#1 Examin#1 Thesi#1	PM Dr Naomie Salim	1
195	Analysi#1 Hierarch#1 Fuzzi#1 Cluster#1 Suggest#1 Supervisor#1 Examin#1 Thesi#1 Tidl#1	PM Dr. Naomie Salim	1
196	Develop#1 Student#1 Perform#1 Evaluat#1 System#1	Dr. Azizah Abd. Rahman	8
197	Develop#1 Custom#1 Relationship#1 Manag#1 System#1 Support#1 Tool#1 Improve#1 Servic#1 Perpustakaan#1 Sultanah#1 Zanariah#1	Dr. Azizah Abdul Rahman	5
198	K#1 Portal#1 Zakat#1	Dr Othman Ibrahim	7
199	Redesign#1 project#1 monitor#1 process:#1 Case#1 studi#1 Pejabat#1 Harta#1 Bina#1 UTM#1	Dr. Azizah Abd. Rahman. Associate Prof. Dr. Rose Alinda Alias	5
200	Comparison#1 Retriev#1 Scheme#2 Base#1 Tidl#1 Abstract#1 Bibliographi#1 Structur#1 Thesi#1 Weight#1	PM Dr Naomie Salim	1
201	Optimiz#1 Process#1 Numer#1 Control#1 Code#1 Manufactur#1 Endmill#1 Tool#1 Endpoint#1 Size#1 20mm#1	Dr Habibollah bin Haron	6
202	Optimiz#1 Numer#1 Control#1 Code#1 Manufactur#1 Ball#1 End#1 25mm#1 Tool#1	Dr Habibollah bin Haron	6
203	Algorithm#1 Enhancement#1 Host#1 Base#1 Intrusion#1 Detect#1 System#1 Discrimin#1 Analysi#1	Prof Dr Abdul Hanan bin Abdullah	2
204	Develop#1 Graphic#1 User#1 Interfac#1 GUI#1 Firewal#1 Monitor#1 System#1	Prof Dr Abdul Hanan bin Abdullah	2
205	Steganographi#1 Cryptographi#1 Apply#1 Hide#1 X-Rai#1 Image#1	Prof Dr Aizaini Maarof	2
206	Improve#1 Two-Term#1 Backpropag#1 Error#1 Function#1 GA#1 Base#1 Paramet#1 Tune#1 Classif#1 Problem#1	PM Dr. Siti Mariyam Hj Shamsuddin	1

APPENDIX F
Supervisor Code

Supervisor Code

CODE	NAME OF SUPERVISOR
1	Rose Alinda Alias
2	Mohd Noor Md Sap
3	Ahmad Zaki Abu Bakar
4	Mohd Aizaini Maarof
5	Shamsul Shahibuddin
6	Abd Manan Ahmad
7	Safaai Deris
8	Zamri Mohamed
9	Harihodin Selamat
10	Abd Hanan Abdullah
11	Abd Samad Hj Ismail
12	Safie Mat Yatim
13	Abd Rahman Ahmad
14	Siti Mariyam Hj Shamsuddin
15	Daut Daman
16	Ghazali Sulong
17	Mohd Salihin Ngadiman
18	Zulkifli Mohamad
19	Noraniah Mohd Yassin
20	Azizah Abdul Rahman
21	Muhammad Shafie Hj Abd Latiff
22	Naomie Salim
23	Ali Selamat
24	Habibollah Haron
25	Othman Ibrahim

Expert Code

CODE	STREAMLINE
1	Data Mining (mdnoor, naomie, safaai, manan, mariyam, noraniah, ali, salihin)
2	Security (hanan, aizaini, kamarul, norbik, zailani)
3	Graphics (mariyam, daut, ghazali, zulkifli, habib, sarudin, safieY, shafieL)
4	Network & Collaborative (samad, shamsul, shafieL, kamarul, asri)
5	Knowledge Management (rose, zaki, shamsul, noraniah, azizah, othman, zamri)
6	Manufacturing (ab, salihin, habib)
7	ISP (rose, wardah, azizah, afida, othman, harihodin, zamri)
8	E-Learning (norazah, ab, naomie, safieY, noraniah, mdnoor)

APPENDIX G

Expert Code

Expert Code

CODE	STREAMLINE
1	Data Mining (mdnoor, naomie, safaai, manan, mariyam, noraniah, ali, salihin)
2	Security (hanan, aizaini, kamarul, norbik, zailani)
3	Graphics (mariyam, daut, ghazali, zulkifli, habib, sarudin, safieY, shafieL)
4	Network & Collaborative (samad, shamsul, shafieL, kamarul, asri)
5	Knowledge Management (rose, zaki, shamsul, noraniah, azizah, othman, zamri)
6	Manufacturing (ab, salihin, habib)
7	ISP (rose, wardah, azizah, afida, othman, harihodin, zamri)
8	E-Learning (norazah, ab, naomie, safieY, noraniah, mdnoor)

APPENDIX H
Ward's Performance

Table 4.2 Ward' result - Sample 50:50

#Doc	ExpertSurvey	Predict (W)
1	5	5
2	1	1
3	5	1
4	2	5
5	2	3
6	5	1
7	1	5
8	1	1
9	2	1
10	5	5
11	5	5
12	2	5
13	5	5
14	5	5
15	5	5
16	5	2
17	7	2
18	4	2
19	5	3
20	1	3
21	1	5
22	2	5
23	4	3
24	5	5
25	7	1
26	1	3
27	5	1
28	8	1
29	4	1
30	3	1
31	5	5
32	1	2
33	7	7
34	5	5
35	5	5
36	8	2

37	1	1
38	1	1
39	7	3
40	3	5
41	3	1
42	3	1
43	1	1
44	5	5
45	1	1
46	1	1
47	3	3
48	8	5
49	5	5
50	1	1
51	5	5
52	5	5
53	1	7
54	5	5
55	5	1
56	5	5
57	5	5
58	1	1
59	2	1
60	1	1
61	1	1
62	5	5
63	3	5
64	2	1
65	5	5
66	1	1
67	5	5
68	4	5
69	1	5
70	1	1
71	1	1
72	5	5
73	5	5
74	6	1
75	2	?

76	3	5
77	5	5
78	1	1
79	5	5
80	5	?
81	8	1
82	5	2
83	5	5
84	5	1
85	8	5
86	5	1
87	5	5
88	7	5
89	5	5
90	1	1
91	1	5
92	1	?
93	3	1
94	3	1
95	2	5
96	5	5
97	5	3
98	3	1
99	5	1
100	1	1
101	2	3
102	2	2
103	5	1
		45.63%

Table 4.3 Ward's Result - Sample
60:40

#Doc	ExpertSurvey	Predict(W)
1	5	5
2	1	1
3	5	1
4	2	5
5	2	3
6	5	1
7	1	5
8	1	1
9	2	1
10	5	5
11	5	5
12	2	5
13	5	5
14	5	5
15	5	5
16	5	2
17	7	2
18	4	2
19	5	3
20	1	3
21	1	5
22	2	5
23	4	3
24	5	5
25	7	1
26	1	3
27	5	1
28	8	1
29	4	1
30	3	1
31	5	5
32	1	2
33	7	7
34	5	5
35	5	5
36	8	2

37	1	1
38	1	1
39	7	3
40	3	5
41	3	1
42	3	1
43	1	1
44	5	5
45	1	1
46	1	1
47	3	3
48	8	5
49	5	5
50	1	1
51	5	5
52	5	5
53	1	7
54	5	5
55	5	1
56	5	5
57	5	5
58	1	1
59	2	1
60	1	1
61	1	1
62	5	5
63	3	5
64	2	1
65	5	5
66	1	1
67	5	5
68	4	5
69	1	5
70	1	1
71	1	1
72	5	5
73	5	5
74	6	1
75	2	?

76	3	5
77	5	5
78	1	1
79	5	5
80	5	?
81	8	1
82	5	2
		48.78%

Table 4.4 Ward's Result - Sample 75:25

#Doc	ExpertSurvey	Predict (W)
20	1	3
21	1	5
22	2	5
23	4	3
24	5	1
25	7	5
26	1	3
27	5	5
28	8	5
29	4	5
30	3	1
31	5	1
32	1	3
33	7	2
34	5	5
35	5	5
36	8	1
37	1	5
38	1	1
39	7	3
40	3	5
41	3	1
42	3	1
43	1	1
44	5	5
45	1	1
46	1	1
47	3	3
48	8	5
49	5	5
50	1	1
51	5	5
52	5	5
53	1	7
54	5	1
55	5	5
56	5	5

57	5	5
58	1	1
59	2	1
60	1	1
61	1	1
62	5	5
63	3	5
64	2	5
65	5	5
66	1	1
67	5	5
68	4	5
69	1	5
70	1	1
		45.10%

Table 4.5 Ward's Result - Sample 80:20

#Doc	ExpertSurvey	Predict(W)
1	5	5
2	1	1
3	5	1
4	2	5
5	2	3
6	5	1
7	1	5
8	1	1
9	2	1
10	5	5
11	5	5
12	2	5
13	5	5
14	5	5
15	5	5
16	5	2
17	7	2
18	4	2
19	5	3
20	1	3
21	1	5
22	2	5
23	4	3
24	5	5
25	7	1
26	1	3
27	5	1
28	8	1
29	4	1
30	3	1
31	5	5
32	1	2
33	7	7
34	5	5
35	5	5
36	8	2

37	1	1
38	1	1
39	7	3
40	3	5
41	3	1
		36.59%

Table 4.6 Ward's Result - Sample 95:5

#Doc	ExpertSurvey	Predict(W)
148	1	1
149	1	1
150	5	5
151	8	5
152	5	5
153	1	1
154	5	5
155	1	1
156	2	5
157	3	1
		70.00%

APPENDIX I
Kohonen Performance

Table 4.8 Kohonen Result - Sample
50:50

#Doc	ExpertSurvey	Predict (K)
1	5	5
2	1	7
3	5	5
4	2	1
5	2	5
6	5	5
7	1	5
8	1	5
9	2	1
10	5	6
11	5	5
12	2	7
13	5	5
14	5	5
15	5	3
16	5	5
17	7	5
18	4	1
19	5	5
20	1	5
21	1	1
22	2	5
23	4	5
24	5	5
25	7	7
26	1	5
27	5	5
28	8	3
29	4	1
30	3	7
31	5	5
32	1	5
33	7	7
34	5	5
35	5	5
36	8	5

37	1	5
38	1	5
39	7	7
40	3	7
41	3	7
42	3	5
43	1	5
44	5	5
45	1	1
46	1	7
47	3	7
48	8	1
49	5	5
50	1	1
51	5	5
52	5	5
53	1	7
54	5	5
55	5	5
56	5	5
57	5	5
58	1	5
59	2	5
60	1	7
61	1	5
62	5	5
63	3	1
64	2	5
65	5	5
66	1	7
67	5	5
68	4	0
69	1	5
70	1	7
71	1	1
72	5	5
73	5	5
74	6	5
75	2	0

76	3	5
77	5	5
78	1	7
79	5	5
80	5	5
81	8	5
82	5	3
83	5	5
84	5	5
85	8	0
86	5	5
87	5	5
88	7	1
89	5	5
90	1	5
91	1	7
92	1	7
93	3	5
94	3	7
95	2	5
96	5	5
97	5	5
98	3	1
99	5	5
100	1	5
101	2	5
102	2	7
103	5	5
		43.69%

Table 4.9 Kohonen Result - Sample 60:40

#Doc	ExpertSurvey	Predict(K)
1	5	1
2	1	1
3	5	5
4	2	5
5	2	5
6	5	5
7	1	1
8	1	1
9	2	5
10	5	8
11	5	1
12	2	5
13	5	5
14	5	3
15	5	1
16	5	8
17	7	7
18	4	1
19	5	5
20	1	1
21	1	5
22	2	8
23	4	1
24	5	5
25	7	1
26	1	1
27	5	5
28	8	3
29	4	1
30	3	5
31	5	1
32	1	1
33	7	7
34	5	1
35	5	1
36	8	8

37	1	6
38	1	5
39	7	7
40	3	3
41	3	5
42	3	3
43	1	1
44	5	5
45	1	5
46	1	5
47	3	3
48	8	8
49	5	5
50	1	5
51	5	5
52	5	1
53	1	7
54	5	5
55	5	1
56	5	5
57	5	5
58	1	3
59	2	3
60	1	5
61	1	1
62	5	5
63	3	3
64	2	7
65	5	5
66	1	5
67	5	5
68	4	3
69	1	1
70	1	5
71	1	5
72	5	5
73	5	8
74	6	7
75	2	8

76	3	1
77	5	1
78	1	7
79	5	5
80	5	8
81	8	7
82	5	1
		42.68%

Table 4.10 Kohonen Result - Sample 75:25

#Doc	ExpertSurvey	Predict (K)
20	1	1
21	1	2
22	2	1
23	4	4
24	5	5
25	7	5
26	1	5
27	5	5
28	8	5
29	4	2
30	3	5
31	5	5
32	1	2
33	7	7
34	5	5
35	5	2
36	8	3
37	1	5
38	1	7
39	7	0
40	3	2
41	3	5
42	3	2
43	1	0
44	5	4
45	1	5
46	1	6
47	3	4
48	8	1
49	5	5
50	1	2
51	5	5
52	5	5
53	1	1
54	5	5
55	5	5

56	5	1
57	5	5
58	1	2
59	2	7
60	1	5
61	1	1
62	5	1
63	3	5
64	2	5
65	5	5
66	1	7
67	5	5
68	4	5
69	1	1
70	1	6
		35.29%

Table 4.11 Kohonen Result - Sample 80:20

#Doc	ExpertSurvey	Predict(K)
1	5	5
2	1	5
3	5	5
4	2	5
5	2	5
6	5	1
7	1	1
8	1	7
9	2	2
10	5	5
11	5	5
12	2	3
13	5	2
14	5	2
15	5	5
16	5	5
17	7	1
18	4	1
19	5	1
20	1	1
21	1	1
22	2	5
23	4	1
24	5	2
25	7	1
26	1	1
27	5	5
28	8	8
29	4	2
30	3	1
31	5	5
32	1	1
33	7	2
34	5	5
35	5	5

36	8	3
37	1	1
38	1	1
39	7	3
40	3	5
41	3	5
		46.34%

Table 4.6 Kohonen Result - Sample 95:5

#Doc	ExpertSurvey	Predict(K)
148	1	1
149	1	7
150	5	1
151	8	8
152	5	1
153	1	1
154	5	5
155	1	1
156	2	2
157	3	2
		50.00%