STATUS OF THESIS

Title of thesis	USING LATENT SEMANTIC INDEXING FOR DOCUMENT CLUSTERING

I, LAILIL MUFLIKHAH

hereby allow my thesis to be placed at the Information Resource Center (IRC) of Universiti Teknologi PETRONAS (UTP) with the following conditions:

1. The thesis becomes the property of UTP
2. The IRC of UTP may make copies of the thesis for academic purposes only.
3. This thesis is classified as

☐ Confidential

☑ Non-confidential

If this thesis is confidential, please state the reason:
_____
_____
_____

The contents of the thesis will remain confidential for _____ years.

Remarks on disclosure:
_____
_____
_____

Endorsed by

_____ Signature of Author

Permanent address:
Perum Puncak Permata Sekaling
Blok N-23, Malang
Jawa Timur, Indonesia

Date : 13 May 2010

_____ Signature of Supervisor

Dr. Baharum Baharudin

Date : 13 May 2010

UNIVERSITI TEKNOLOGI PETRONAS

DISSERTATION TITTLE: USING LATENT SEMANTIC INDEXING FOR DOCUMENT CLUSTERING

by

LAILIL MUFLIKHAH

The undersigned certify that they have read, and recommend to the Postgraduate Studies Programme for acceptance this thesis for the fulfillment of the requirements for the degree stated.

Signature: _____

Main Supervisor: Dr. Baharum Baharudin

Signature: _____

Co-Supervisor: _____

Signature: _____

Head of Department: Dr. Mohd Fadzil Hassan

Date: 13 May 2010

USING LATENT SEMANTIC INDEXING FOR DOCUMENT CLUSTERING

by

LAILIL MUFLIKHAH

A Thesis

Submitted to the Postgraduate Studies Programme

as a Requirement for the Degree of

MASTER OF SCIENCE

COMPUTER AND INFORMATION SCIENCE DEPARTMENT

UNIVERSITI TEKNOLOGI PETRONAS

BANDAR SERI ISKANDAR,

PERAK

MAY 2010

DECLARATION OF THESIS

Title of thesis	USING LATENT SEMANTIC INDEXING FOR DOCUMENT CLUSTERING

I, LAILIL MUFLIKHAH

hereby declare that the thesis is based on my original work except for quotations and citations which have been duly acknowledged. I also declare that it has not been previously or concurrently submitted for any other degree at UTP or other institutions.

Witnessed by

_____ _____

Signature of Author

Permanent address:
Perumahan Puncak Permata Sengkaling
Blok N-23, Malang, Jawa Timur
Jawa Timur, Indonesia

Date : 13 May 2010

Signature of Supervisor

Dr. Baharum Baharudin

Date : 13 May 2010

DEDICATION

This thesis is dedicated to my mother who always tirelessly support and pray for me. For all my sisters, my brother and my teachers on every grudge and every fight you are my everlasting inspiration and I miss you all day and night. There is light on every night, there is hope on every fright. I hope this thesis gives a contribution for a better life.

ACKNOWLEDGEMENTS

I would like to thank Allah subhanahu wata'alaa, Muhammad shallahu 'alaihi wasallam the most inspiration of my life. This work could not have been possible without the advice and support of so many people. First and foremost, I would like to thank my research supervisor the honorable Dr. Baharum Baharudin for his nice and full-dedicated guidance during conducting my research project. I would like to thank Dr. Mohd Fadzil Hassan for their helpful comments, comprehension, and reviewing my thesis. Also, I would like to thank Universiti Teknologi PETRONAS for providing well-equipped facilities to conduct this research.

I am also deeply grateful to my mother and my sisters, whose support went well beyond kindly words of encouragement and finally become for me an affirmation that I really did have something worthwhile to contribute for others. I am deeply thankful to my family and friends who support me to pursue my master studies. Also to everyone who has given contribution to this thesis directly or indirectly I really appreciate it.

ABSTRACT

Documents with various contents are easily obtained from URLs which are associated with their titles. However, the titles of documents may not describe their contents and they just attract the readers to buy and read them. Therefore, the document clustering based on the same category is important to help users to retrieve information they need. Document clustering is an implementation of data mining task. By using similarity measurement of documents' characteristic, they can be clustered based on the same category or topic. High dimensionality of the document representation is due to representing of all substantial words in the vector space model. It is one of problems in document clustering that decreases the cluster quality performance including f-measure, entropy and accuracy. In categorical domain, many research have been conducted to reduce the dimension size of term-document matrix representation until by using keyword base. However, the result is obtained low accuracy in various class sizes of document collections. Therefore, this research is intended to improve the quality and accuracy of document clustering by using a method in information retrieval.

A method in information retrieval, Latent Semantic Indexing (LSI), is proposed to reduce the dimension of term-document matrix for document representation. In this work, the LSI method is used to produce the patterns of terms, so that documents can be mapped into concept space. Based on the new representation, the documents are then subjected to the clustering algorithm itself, which is Fuzzy c-Means algorithm. A variant of distance measurement, cosine similarity, is also embedded to this algorithm. The results are then compared with some existing algorithms, which are used for benchmark purposes. The results show that the proposed method obtains high quality cluster and it is superior to the other fuzzy clustering algorithms for category i.e. FCCM, FSKWIC, and Fuzzy CoDoK with accuracy rate of over 90%.

ABSTRAK

Dokumen yang mempunyai banyak maklumat dan isi kandugan amat mudah diperoleh dari internet dengan mengenal pasti tajuk dokumen. Tetapi, tajuk dokumen tersebut tidak selalunya menggambarkan isi kandungan dokumen tersebut. Namun disebabkan tajuk dokumen itu, ramai individu akan tertarik untuk membeli dan membacanya. Disebabkan faktor ini, membahagikan dokumen ke dalam kategori yang sama dan tepat amat penting bagi membantu individu untuk memperoleh maklumat yang mereka inginkan. Pembahagian dokumen ialah satu cara kita dapat mengaplikasikan penyelidikan maklumat. Dengan menggunakan ukuran persamaan pada ciri-ciri dokumen tersebut, ia boleh dibahagikan kepada kategori atau tajuk yang sama. Dokumen yang terperinci dan mempunyai dimensi yang tinggi akan mempamerkan semua perkataan substansial di dalam model ruang vektor, dan ini merupakan satu masalah dalam pembahagian dokumen kerana ia boleh mengurangkan kualiti keupayaan pembahagian dokumen seperti f-measure, entropy dan juga ketepatan. Banyak kajian telah dijalankan pada categorical-domain dengan menggunakan kata kunci dengan tujuan untuk mengurangkan saiz dimensi dari pemaparan matrik terma-dokumen. Namun demikian, bila ia di aplikasikan kepada pelbagai jenis dan saiz dokumen, ketepatannya akan berkurangan. Oleh kerana itu, kajian ini bertujuan untuk meningkatkan kualiti dan ketepatan pembahagian dokumen kepada kategori yang sama dengan menggunakan kaedah penyelidikan maklumat.

Latent Semantic Indexing (LSI) adalah satu kaedah bagi memperoleh maklumat dan kaedah ini disarankan untuk mengurangkan pemaparan matrik terma-dokumen. Dalam kajian ini, kaedah LSI digunakan untuk menghasilkan pola ayat-ayat supaya dokumen-dokumen dapat dipetakkan dalam ruangkonsep. Berdasarkan kaedah pengolahan baru ini, dokumen akan disubjekkan pada algoritma pembahagian Fuzzy c-Means. Algoritma ini turut menggunakan pelbagai kaedah ukuran jarak, Cosine Similarity. Hasil kajian tersebut akan dibandingkan dengan algoritma yang lain dan

akan dijadikan sebagai kayu pengukur. Keputusan yang diperoleh dengan menggunakan kaedah ini menunjukkan bahawa kaedah ini adalah lebih baik dan cemerlang berbanding FCCM, FSKWIC, dan Fuzzy CoDoK dengan purata ketepatannya melebihi 90%.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS vi
ABSTRACT vii
TABLE OF CONTENTS x
LIST OF FIGURES xii
LIST OF TABLES xiii
CHAPTER 1 INTRODUCTION 1
1.1 Background 1
1.2 Motivation 3
1.3 Problem Statement 4
1.4 Objective 4
1.5 Scope of Study 4
1.6 Thesis Contribution 5
1.7 The Outline of Thesis 5
CHAPTER 2 RELATED WORK 6
2.1 Introduction 6
2.2 Clustering System 6
2.2.1 Pattern Representation, Feature Selection and Feature Extraction 6
2.2.2 Pattern Proximity Measure (Inter pattern Similarity) 7
2.2.3 Clustering or Grouping 7
2.3 Hierarchical versus Non-Hierarchical Clustering 9
2.4 Exclusive (Hard) versus Overlapping versus Fuzzy Clustering 11
2.5 Document Clustering 13
2.5.1 Fuzzy Document Clustering for Document Categorization Based on Key-Word 14
2.5.2 Fuzzy Document Clustering for Document Categorization Using Ontological Approach 15
2.6 Summary 16
CHAPTER 3 THEORETICAL BACKGROUND 18
3.1 Introduction 18
3.2 Information Retrieval 18
3.2.1 Latent Semantic Indexing 20
3.2.1.1 Singular Value Decomposition (SVD) 21
3.2.1.2 Principal Component Analysis (PCA) 22
3.2.2 Applying Latent Semantic Indexing into Information Retrieval 24
3.3 Fuzzy Clustering 28
3.3.1 Fuzzy c-Means Clustering 31

3.3.2 Similarity Measurement in Fuzzy c-Means 33
3.3.3 Fuzzy Prediction Methods 36
3.4 Summary 38
CHAPTER 4 METHODOLOGY 39
4.1 Introduction 39
4.2 Preprocessing Step for Document Clustering 40
4.3 Representation of Document Collection 44
4.4 Dimension Reduction of Term-Document Matrix Representation 46
4.5 Mapping Term-Document Matrix to Document Matrix Using LSI Approach 48
4.6 Modified Fuzzy c-Means Algorithm 48
4.7 Summary 51
CHAPTER 5 EXPERIMENTAL RESULT AND DISCUSSION 52
5.2 Data Preparation 52
5.3 Result of Fuzzy c-Means Clustering Algorithm 57
5.3 Evaluation Measurement for Document Clustering 60
5.3.1 Entropy 60
5.3.2 F-Measure 60
5.4 Performance Evaluation of Document Clustering 62
5.4.1 Confusion Matrix 67
5.4.2 Comparison of accuracy to FCCM, FSKWIC, and Fuzzy CoDoK ... 70
5.5 Summary 71
CHAPTER 6 CONCLUSSION AND FUTURE WORK 72
6.1 Conclusion 72
6.2 Future Work 74
REFERENCES REFERENCES 75
LIST OF PUBLICATIONS 80
APPENDIX A Stop Word List 81
APPENDIX B Document Form 83
APPENDIX C Performance of Document Clustering with Various *K*-Rank 86
APPENDIX D Result Illustration of Fuzzy c-Means Clustering Algorithm 88
APPENDIX E Document Clustering Result 91

LIST OF FIGURES

Figure 1.1 Clustering system 3
Figure 2.1 Different way of representing clusters 8
Figure 2.2 Simple classification of clustering methods 9
Figure 3.1 (a) Using concept for retrieval and (b) similarity ranking of document 19
Figure 3.2 Getting pattern in data collection 20
Figure 3.3 Reduced dimension representation of term-document matrix 21
Figure 3.4 Document and query vector in similarity concept 28
Figure 3.5 Membership function of fuzzy set 29
Figure 3.6 Geometric illustration of the Cosine measure 34
Figure 3.7 Properties of (a) Euclidean-based, (b) Cosine similarity illustrated in two dimension. 35
Figure 4.2 Stemming process using Porter's algorithm 43
Figure 4.3 Flow chart of modified Fuzzy c-Means clustering 50
Figure 5.1 An original document from data set 'Classic3' 54
Figure 5.2 A document 'Classic3' after *preprocessing* 55
Figure 5.3 Illustration of Fuzzy c-Means clustering algorithm by the data set 'Binary2' 58
Figure 5.4 Performance evaluation of clustering 'Multi5' with various *k*-ranks 62
Figure 5.5 Performance evaluation of clustering 'Multi5'with various *k*-ranks 63
Figure 5.6 Class size versus number of *k*-rank 64
Figure 5.7 Cluster number versus number of *k*-rank 64
Figure 5.8 Comparison F-measure for all data sets 66
Figure 5.9 Entropy of document clustering 67

LIST OF TABLES

Table 2.1 Various types of clustering methods....8
Table 2.2 Complexity of clustering algorithms (Jain 1988) 11
Table 4.1 Weight term representation of document collection....46
Table 5.2 Data of thresholds....55
Table 5.3 Data reduction details after *preprocessing* 56
Table 5.4 Vector Space Model....57
Table 5.5 The number of document clustering result 59
Table 5.6 Optimum *k*-rank of data sets....63
Table 5.7 Precision of document clustering....65
Table 5.8 Recall of document clustering 65
Table 5.9 F-measure of document clustering for all data sets 66
Table 5.10 The data details of entropy of document clustering for all data sets 67
Table 5.11 Confusion Matrix for data set 'Binary2' using SVD....68
Table 5.12 Confusion Matrix for data set 'Multi5' using SVD....68
Table 5.13 Confusion Matrix for data set 'Multi10' using SVD....69
Table 5.14 Confusion Matrix for data set 'Classic3' using SVD 69
Table 5.15 Confusion Matrix for data set 'YahooK1' using SVD 70
Table 5.16 The details of misclassified document clustering using PCA 70
Table 5.17 Comparison of accuracy using different algorithms....71

CHAPTER 1

INTRODUCTION

1.1 Background

In recent years, the number of documents on Internet has been growing vastly. There are many topics related to education, finance, sports, entertainment, etc. It is important to know the topics of the document before getting more information. Data mining is often used as a technique for finding meaningful patterns from large data collection. This technique aims to find and describe structural patterns in data collection as a tool for helping researchers to explain the data and make predictions from it.

Generally, data mining tasks are divided into two major categories: *predictive* and *descriptive* task. The purpose of the *predictive* task is to predict the value of a specific attribute based on the values of other attributes. The predicted attribute is known as the dependent variable (objective). Another attribute that is used for making prediction is known as the independent variable (explanatory). The second category aims to get patterns (correlations, trends, clusters, trajectories, and anomalies) that analyze the main points of the essential correlations in the data [1]. As a *descriptive* task, the data collection can be known whether one is in the same partition (group) or not. Clustering is a method for arranging a large data collection by partitioning the data set, so that objects in the same cluster are more similar to one another than to objects in other clusters. The clustering methods can be classified into *hierarchical method* and *partitional method*. The *hierarchical* method uses an $N \times N$ similar matrix, containing the pair-wise similarities in a dataset of size N objects, to make a nested set of clusters. Meanwhile, the *partitional method* can be divided into *exclusive* (hard clustering) and *overlapping* (fuzzy clustering). In hard clustering, an object can only belong to one cluster, whereby we assume well defined boundary among the clusters.

But, in any given document, there is a possibility that it can consist of multiple subjects or categories. For this, the fuzzy clustering approach is used. Thus, the context of this thesis is implementation of data mining method to cluster document in the similar category.

The clustering of document is different from the clustering of symbolic data, such as chemical, medical, biological, etc. which is grouped based on the similarity of term or symbol as attributes without considering the entire content (meaning). However, by referring to the content, the clustering of document can be used to classify documents based on the same category or topic. Document clustering is related to the arrangement of a large data text collection. Some terms or words inside each document is often irrelevant with the topic or content such as common word, conjunctions, or words used just for introducing to topic (minor words). In contrast, there are correlations of terms among documents. Therefore, the user needs to remove the irrelevant data and replace it with meaningful data in order to cluster the document based on the topic or category on the same concept.

In the field of *information retrieval* (IR), document clustering is used to provide clusters of documents under the same topic, using user's browsing (query) of retrieval result [2]. The implementation of the document clustering has a benefit in IR application as mentioned in [3]. Document clustering has always been used as a tool to improve the performance of retrieval and navigating large data. The document clustering has been used to enhance information retrieval. A collection of text (document) is clustered into groups or clusters having similar contents. In some applications, the clustering is called data segmentation because it involves the clustering partitions of large data sets into groups according to their similarity of the content. Generally, a clustering system is illustrated as shown in Figure 1.1[4].

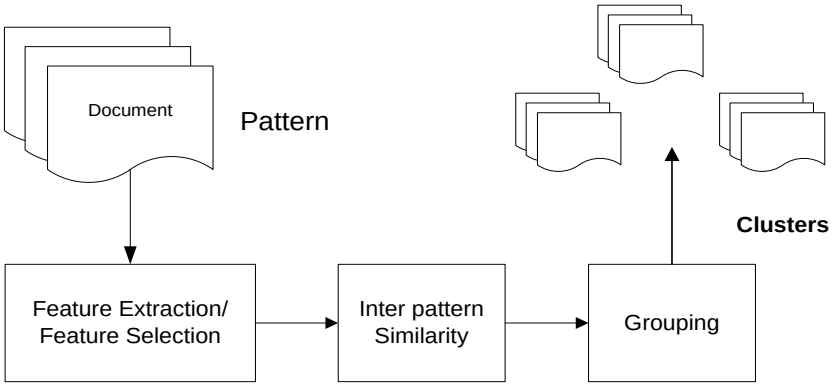


Figure 1.1 Clustering system

Figure 1.1 depicts a clustering system where document features are extracted from the original data (document collection). In the feature extraction, the document undergoes a *preprocessing* step (which includes *parsing*, *stemming*, *removing stop word*) in order to remove the meaningless data. Then, indexing of term-document is carried out to define similarity and dissimilarity between one document and the others in document representation. The similarity measurement is used to group the documents into their respective clusters. Documents with high similarity will be grouped into the same cluster. Otherwise, it will be assigned to the different cluster. However, clustering result uses indexing of weight-terms is still obtained low performance quality. All words or terms with the same frequency in a document collection are defined as the same weight, even though they have no relevance to the category. In other word, the document clustering just based on the morphology.

1.2 Motivation

This research is motivated by the previous experiments done either by researchers or by the author in document clustering based on category similarity [5, 6]. In those works, documents are subjected to the common preprocessing techniques (parsing, stemming, removing stop word) in order to get the significant terms. However, the clustering results did not prove to be encouraging in terms of f-measure, entropy, precision and recall. Therefore, this research proposes to find ways to see if further improvement can be made.

1.3 Problem Statement

High dimensionality of term-document matrix is caused by the process of composing words and terms into document collections. A document is composed of a number of terms (words) such as substantial words, common words, prepositions, articles and numbering as well as many tags that contain meaningless information. Many research have been conducted to reduce the dimension using various algorithms. However, overall performance of the results obtained is low in accuracy and f-measure but high in entropy either for small or large class size. Besides, document clustering based on the frequent word occurrence often gives inaccurate results due to synonymy and polysemy problems. Therefore, one of major issues addressed in this research is to get the correct representation of document.

1.4 Objective

Generally, the main objective of this thesis is to improve the performance either internal or external quality of document clusters in term of entropy, f-measure, and accuracy, by embedding Latent Semantic Indexing approach. The details are follows.

- To apply *preprocessing* document to get the meaningful terms inside the document collections.
- To map document collections into concept space in vector domain to solve the synonymy and polysemy problems
- To select the dissimilarity formula to be involved in the objective function of Fuzzy c-Means clustering algorithm

1.5 Scope of Study

The scope of this thesis is that the document clustering is not considered semantic feature of words explicitly, but it just involves syntax of words. Therefore, it will assume that there is no correlation among the terms or words that have the same meaning.

1.6 Thesis Contribution

The main contributions for document clustering are given below.

- Latent Semantic Indexing method is applied to handle the high dimensionality problem that represent document in concept space within a vector domain. It is also used to solve synonym and polysemy problems.
- Cosine similarity is incorporated in objective function of Fuzzy c-Means clustering algorithm

1.7 The Outline of Thesis

This thesis has six chapters and is organized in the following manner. Chapter Two presents a review of related work on document clustering. This chapter includes the general concept of theory and drawback of the existing algorithms.

Chapter Three consists of descriptions on the two main methods used for this research. The first method is related to the reduction of matrix dimension for document representation by using the Latent Semantic Index approach. The second is on the fuzzy clustering algorithm (Fuzzy c-Means), that will be used to cluster the document collections.

Chapter Four covers the methodology of the research that will be used as a frame work. It starts with document *preprocessing*, representing text document in concept space, and applying the Fuzzy c-Means algorithm for clustering.

Then, Chapter Five discusses the experimental results of the proposed method. It includes analysis and comparison of the clustering performance (precision, recall, and f-measure as well as the entropy).

Finally, the conclusion of this research is put in Chapter Six. In this chapter, we also suggest future work that can be carried out to enhance this research.

CHAPTER 2

RELATED WORK

2.1 Introduction

This chapter consists of related work conducted by many researchers with various approaches to solve document clustering problems in information retrieval. The merits and demerits of some methods will also be highlighted. The chapter also includes clustering system and definition of various methods in clustering which is explained in the beginning part.

2.2 Clustering System

Clustering refers to the task of partitioning unlabelled data into meaningful group. It can be considered as the most important unsupervised learning problem. Clustering can be defined as the process of organizing objects into groups whose members are similar qualities or attributes. The process involves grouping data into several new classes and a common descriptive task whereby one seeks to identify a finite set of categories or clusters to describe the data. The process of clustering activities that affects the output of a clustering system is illustrated by Jain et al.[7, 8]:

1. Pattern representation, optionally including feature extraction and/ or selection
2. Definition of a pattern proximity measure for the data domain
3. Clustering or grouping of data points according to the chosen pattern representation and the proximity measure.

The output of a clustering system is the result of the system's interactive activities of the above stage. Thus, the detail information is described as bellow:

2.2.1 Pattern Representation, Feature Selection and Feature Extraction

Pattern representation refers to feature representation which is a paradigm for observation and abstraction of the learning problem, including the type, the number

and the scale of the features, the number of the patterns, and the format representation. Then, the future selection is defined to identify the most representative subset of the natural features or transformations of the natural features to be used by the machines. The feature set also affects the quality as well as the efficiency of cluster system. A large feature set containing numerous irrelevant features does not improve the clustering quality but increases the computational complexity of the system. On the other hand, an insufficient feature set may decrease the accuracy of the representation.

2.2.2 Pattern Proximity Measure (Inter pattern Similarity)

Pattern proximity evaluates the similarity or dissimilarity between two patterns which is known as relationship between two patterns. This measurement serves as the basis for cluster generation as it indicates how two patterns “look alike” to each other.

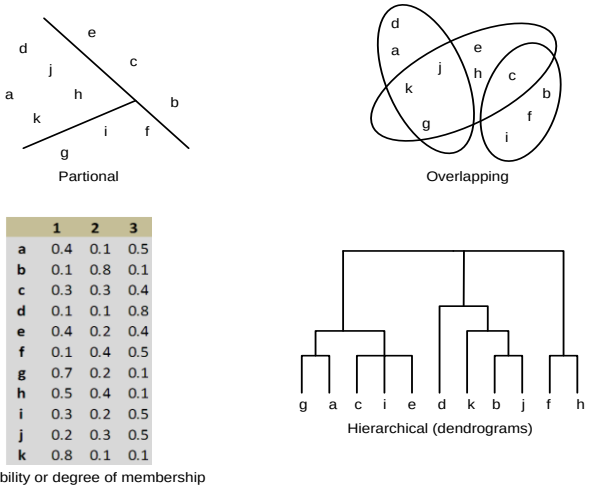
2.2.3 Clustering or Grouping

Data points are clustered or grouped according to the chosen pattern representation and the proximity measure. A clustering algorithm groups the input data according to the set of predefined criteria. The clustering algorithm used by a system can be either statistical or heuristic. There are various types of clustering methods, based on learning paradigm, number of output cluster, cluster assignment, and system architecture respectively as shown in Table 2.1.

Table 2.1 Various types of clustering methods

Criteria	Categories
Learning paradigm	Off-line: iteratively batch learning on the whole input set On-line: incremental learning that does not remember the specific input history
Codebook size (number of output clusters)	Static-sizing: number of cluster is fixed Dynamic-sizing: number of cluster is adaptive to the distribution of input data
Cluster assignment	Hard: Each input data is assigned with one class Fuzzy: Each input data is given a degree of membership with every output cluster
System architecture	Partitioning: input data is separated into disjoint output clusters Hierarchical: the relation of cluster is shown in output tree Density-based: input data are grouped based on density conditions Grid-based: spacial input space is quantized into finite sub-spaces (grids) before clustering of each-sub-space

Data clustering is a process of grouping principal or conceptual objects into classes of similar data objects. There are several ways of representing clusters such as *partitional*, *overlapping*, using *probability* (degree) of membership and *hierarchical clustering* (*dendrograms*) as shown in Figure 2.1 [9].



	1	2	3
a	0.4	0.1	0.5
b	0.1	0.8	0.1
c	0.3	0.3	0.4
d	0.1	0.1	0.8
e	0.4	0.2	0.4
f	0.1	0.4	0.5
g	0.7	0.2	0.1
h	0.5	0.4	0.1
i	0.3	0.2	0.5
j	0.2	0.3	0.5
k	0.8	0.1	0.1

Figure 2.1 Different way of representing clusters

Clustering analysis helps construct meaningful grouping of a large set of objects based on divide-and-conquer methodology that breaks a large-scale system down into

smaller components to simplify design and implementation. Principally, the task of clustering is to maximize the intra-class similarity and minimize the interclass similarity. The main types of cluster analysis methods (system architecture) are the *non hierarchical* clustering which divides a data set of *N* items into *M* clusters and *hierarchical* clustering which produces a nested data set in which pairs of items or clusters are correctly linked. Barnard and Downs classifies the detailed methods of clustering as shown Figure 2.2 [10].

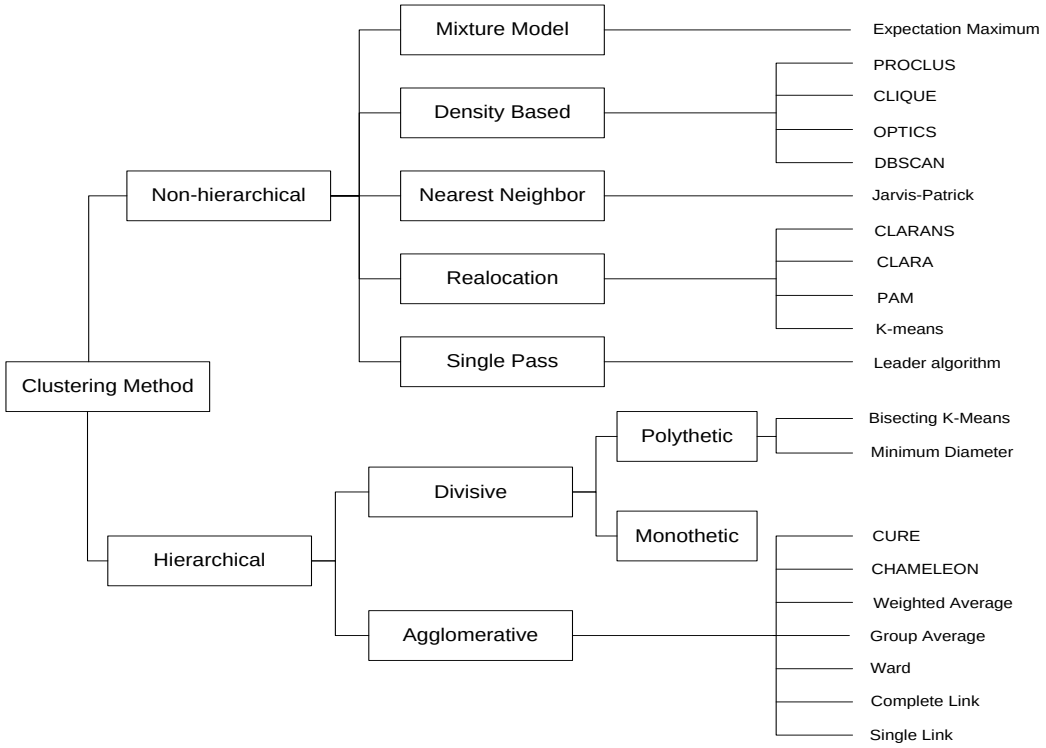


Figure 2.2 Simple classification of clustering methods

2.3 Hierarchical versus Non-Hierarchical Clustering

Hierarchical clustering method is a set of nested clusters organized as a tree. Each node, except the leaves, joins to form sub clusters and the root of tree is the cluster containing all the items. *Non hierarchical* (*partitional)* clustering, on the other hand, is simply a division of a set of data objects taken into non-overlapping subsets (clusters) in which each data object is exactly in one subset.

There are four different types of *hierarchical* clustering: s*ingle link, complete link, group average link,* and w*ard's method.* In the *single link,* each step makes a link for

the most similar pair of the object before performing it in the same cluster. The link can be applied in a relatively efficient manner and it has some interesting methodology properties [3], whereby its cluster structure tends to make suitable chaining for ellipsoidal cluster. In the *complete link*, it uses the minimum similarity in pair between each of two clusters to define the similarity of inter-cluster. All objects in a cluster within some minimum similarity are joined to one another. The characteristic of this clustering type is small and compactly bound clusters. The third type of clustering is *group average link*. In this type, all objects contribute to inter-cluster similarity. This method uses the average values of the pair-wise joins between the defined similarities in a cluster. The last type of clustering is the *Ward's method*. It makes a link on the cluster pair using minimum variance analysis method to evaluate the distances between cluster centers. It has a tendency to create the same kind of clusters and symmetric hierarchy, and its cluster center can be used to represent a cluster. The evaluation of this type shows that it is good when improving the cluster structure although it is sensitive to the outlier and is not good at improving the extended cluster.

The *partitional clustering*, on the other hand, is known as non *hierarchical method* and consists of five types. The first type is *Single Pass*. It is uncomplicated and need to process the data set only once. The second type is *Relocation* which has some initial group of data set and the position of item changes from one cluster to another the grouping improves. The third type is *Nearest Neighbor* whereby each unlabeled pattern is assigned to the nearest cluster labeled as the neighbor pattern. In this type, it is also provided the distance to the labeled neighbor as a threshold. The fourth type is *Mixture Model*. In this method, each cluster is represented by a parametric distribution such as Gaussian for continuous data or Poisson for discrete data and the entire data set is modeled by a mixture of that distribution. Then, the last type is *Density* based clustering. It determines the cluster based on the distribution of generated pattern of high and low density.

There are some different complexities of time and space in the clustering method. This will be relevant when the clustering methods are applied as shown in Table 2.2.

Table 2.2 Complexity of clustering algorithms (Jain 1988)

Clustering Algorithm	Time Complexity	Space Complexity
Leader	$O(kn)$	$O(k)$
K-Means	$O(nkl)$	$O(k)$
ISODATA	$O(nkl)$	$O(k)$
Shortest spanning path	$O(n^2)$	$O(n)$
Single link	$O(n^2 \log n)$	$O(n^2)$
Complete link	$O(n^2 \log n)$	$O(n^2)$

In Table 2.2, the first three algorithms, Leader, K-Means, and ISODATA are classified as *partitional* clustering method, while the rests are classified as different approaches of *hierarchical* clustering method. The number of patterns being clustered is notated by *n*, the number of clusters by *k* and the number of iterations by *l*. This table shows that the time and space complexity of *hierarchical* methods are higher than the *partitional* clustering methods.

2.4 Exclusive (Hard) versus Overlapping versus Fuzzy Clustering

The other types of clustering based on the cluster assignment are exclusive, overlapping, and fuzzy clustering. The exclusive clustering assigns each object to a single cluster. Therefore an object can only belong to one cluster. In overlapping clustering, in contrast, an object can be assigned to more than one cluster. Fuzzy clustering assigns a membership weight between 0 and 1 [0…1] to each object. The fuzzy clustering is similar to probabilistic techniques which compute the probability with which each point belongs to each cluster and the total of probability for each point must equal to one.

Furthermore, there are various types of fuzzy clustering [11]:

1. Fuzzy clustering based on fuzzy relation

A relation between data object is applied to cluster data. The relational format of data assumes that there are degrees of dissimilarity between the patterns and original patterns. The typical predicate or relation can be defined with the concept of similarity or dissimilarity between pairs of patterns. The other predicate may be one of

complementary character, expressing a degree of difference or dissimilarity between pairs of patterns.

2. Fuzzy clustering based on objective function and fuzzy covariance matrix

The clustering concerns with building partitions (cluster) of data sets on the basis of some performance index known as an objective function. The minimization of a certain objective function can be treated as an optimization approach leading to some suboptimal configuration of the clusters.

Furthermore, the objectives function-based fuzzy clustering algorithms include the following methods:

- Fuzzy c-Means algorithm: spherical clusters of approximately the same size
- Gustaftson-Kessel algorithm: ellipsoidal clusters of approximately the same size; there are also axis-parallel variants of this algorithm
- Fuzzy c-varieties algorithm: detection of linear manifolds (infinite lines in 2D)
- Adaptive fuzzy c-varieties algorithm: detection of line segments in 2D data
- Fuzzy c-shells algorithm: detection of circles
- Fuzzy c-rings algorithm: detection of circles
- Fuzzy c-quadric shells algorithm: detection of ellipsoids
- Fuzzy c-rectangular shells algorithm: detection of rectangles

3. Nonparametric classifier, that is the fuzzy generalized *k*-nearest neighbor rule

This method is based on the fuzzy labeling of samples by means of a membership function. As long as the membership function carries more information than the classical characteristic function, the concepts of clustering and discrimination are extended to the fuzzy field. The learning set is fuzzily clustered where by the membership function is based on a weighted *k*-nearest neighbors rule and does not require the optimization of any criterion.

4. Neuro-Fuzzy Clustering

It refers to combinations of neural networks and fuzzy logic in partitioning (clustering). It uses a fuzzy inference system regularly and the tuning of the model parameters is done by neural network learning rules, i.e. Takagi Sugeno (TS) model. In a fuzzy clustering approach, the data space is partitioned directly into clusters and a neuro-fuzzy model is identified from their parameters. There are various algorithms of this type as below:

- Self Organizing Maps
- Fuzzy Learning Vector Quantization
- Fuzzy Adaptive Resonance Theory
- Growing Neural Gas
- Fully Self-Organizing Simplified Adaptive Resonance theory
- Fuzzy Competitive Learning

2.5 Document Clustering

Document clustering is applied largely in many fields, including text mining and information retrieval. Initially, document clustering has been studied to increase the performance (precision and recall) of information retrieval systems [12, 13]. Recently, clustering has been proposed to browse document collection [14] or to organize the results returned by a search engine in response to a user's query [15]. Document clustering has also been used to generate document clusters hierarchically such as the taxonomy of Web documents provided by search engines.

Document clustering mainly used to group similar texts together. There are four basic steps in document clustering process:

1. Generate document representation using Vector Space Model.
2. Select an appropriate similarity measure
3. Use an appropriate clustering method to cluster the document collection
4. Analyze the result

In step 3, there are various methods of clustering, either hierarchical or non-hierarchical (partitial) clustering algorithm itself. Many research have been conducted for document clustering using hard clustering methods [14, 16-18]. They have shown that the performance of *Bisection K-means* (variant of *Basic K-means*) is better than *Agglomerative Hierarchical* clustering. Two techniques of document clustering used widely are *Agglomerative Hierarchical* clustering and *K-means*. The *Agglomerative Hierarchical* clustering is superior to *K-means* for non-document data [19]. Cutting et al. used the hybrid of *K-means* and *Agglomerative Hierarchical* clustering in a document browsing [14]. The *K-means* is used for its efficiency and *Agglomerative*

Hierarchical clustering for its quality. In document hierarchies, *agglomerative* clustering is better than *K-means*.

In the real application, the fuzzy clustering is often transformed to hard clustering by assigning each object to the cluster in which its membership weight or probability is the highest. *Fuzzy c-Means* clustering algorithm has been applied to symbolic data, such as biological, chemical, etc and the result shows that the quality clustering is better than *hierarchical* method (*hard* clustering) [20]. It is also applied to text mining [21]. Recently, document clustering is applied for categorization by using fuzzy clustering approach. However, the result performance in either internal quality (entropy) or external quality (f-measure and accuracy) evaluation is still unsatisfactory, particularly in large data volume [5]. Therefore, reducing data volume, either using keyword or ontological approach, has become the focus of recent research in document clustering.

2.5.1 Fuzzy Document Clustering for Document Categorization Based on Key-Word

In fuzzy domain for categorical data, Oh et al [22] introduces *Fuzzy Clustering for Categorical Multivariate Data* (FCCM). It is a clustering algorithm based on minimization to the sum of intra-cluster distances (or the sum of inter-point distances within clusters in the case of relational data). The attributes can be categorical (nominal) and the distance or similarity between two patterns is not explicitly available as the objective of this algorithm. However, the weakness of this algorithm occurs when it is applied to a large value of data and categorical attribute, especially in text data represented by words. It can lead to numerical instabilities due to overflows.

Another algorithm interpreted as a fuzzy co-clustering algorithm is *Fuzzy Simultaneous Key Word Identification and Clustering of text documents* (FSKWIC)[23]. This algorithm obtains optimum cluster by minimizing the aggregate of the weight distances between object (keyword) and object clusters' centers using cosine similarity or Jackard index to evaluate the similarity or dissimilarity between the documents. The reason for applying this algorithm is the high dimensionality of

the data and the fact that two documents may not be considered similar if keywords are missing in both documents. However, the weight of keyword representative is considered in a cluster even though it not explicitly stated.

Another clustering algorithm, the *Spherical K-means* (SKM) [24] essentially maximizes the sum of intra-cluster similarities using the cosine measure. This algorithm is applied to modify objective function in Fuzzy c-Means known as *Spherical Fuzzy c-Means* (SFCM). *Fuzzy Co-Clustering of Documents and Keywords* (Fuzzy CoDoK) is introduced by Kummamuru et al.[25]. It covers the FCCM algorithm for large text corpus by modifying the membership and centroid function using the Gini regularization index. When it is applied, the performance of FCCM is poor compared to the other analysis of the objective function. The main differences mainly come from constrain and regulation terms used within it. Generally, the Fuzzy CoDoK is observed to perform better than other co-clustering algorithms. The SKM and SFCM performed better than Fuzzy CoDoK when the dataset used consists of a small number of distinct clusters.

2.5.2 Fuzzy Document Clustering for Document Categorization Using Ontological Approach

Many studies try to use ontological based approach to help information retrieval and text document processing. The ontology based method is divided into two types. The first is applied in machine learning methods, such as clustering analysis and fuzzy logic. Both of them are used to build text document and to assist information retrieval and text document processing [26-30]. However, the drawback of this method is that its performance is dependent on a good ontology construction. This method also assumes that the terms which are rarely used in text document are called meaningless. However, it is different from the information theory that the terms rarely used have important thing in information retrieval. The second type is the existing ontology (WordNet). It is used to help information retrieval and text document processing. Three approaches are proposed for this type. The first approach is using WordNet to solve synonyms or hypernyms by replacing them with the keywords. For example, by using hypernym-based methods, the word ‘meat’ can be used to replace ‘beef’ and ‘pork’. The second approach is using word sense disambiguation to solve problem of

synonym and polysemy in natural language. Finally, the third approach is applying various techniques [31-34] to find out the semantic similarities and correlations of terms to enhance the keywords-based information retrieval and text document processing methods. These methods have drawback to the noise initiated by incorrect senses retrieved from WordNet. They are concluded by part of speech (PoS) on disambiguation with lack of knowledge background in information retrieval and text document processing. The PoS tagging method is introduced by Simon [30] to solve this problem by proposing a document search technique. This method is intended to cluster the search results into meaningful categories based on the words that change the original search in the text document.

In general, clustering based on ontology does not give good results. Many features (words or terms) are replaced to the same concept even though in the different content. Misclassifications often happen because of ignoring the impact of the semantic similarities and the relationships among different terms in the same text document in information retrieval and text document preprocessing.

Since the results of above approaches, key-word and ontology to reduce dimension size of data volume, are still unsatisfactory, this research applies another method to generate patterns from the document collection. The method is found in information retrieval known as Latent Semantic Indexing. By embedding this method for representing document, it is proposed to improve the performance quality of document clustering.

2.6 Summary

Clustering system including pattern representation, pattern proximity, and clustering method of data points is described. There are various types of clustering representation. Based on the system architecture, clustering methods can be categorized as *hierarchical* and *non-hierarchical* clustering. Generally, the *non-hierarchical* clustering methods are better than *hierarchical* clustering methods in term of time complexity. Based on the cluster assignment, clustering is classified as hard and fuzzy clustering. The result of fuzzy clustering approach is better than hard clustering approach. However, in fuzzy domain, clustering for categorical data using

keyword and ontology approach, still results in unsatisfactory performance. Thus, this research uses another method, i.e. Latent Semantic Indexing in order to increase the performance quality of clustering result.

CHAPTER 3

THEORETICAL BACKGROUND

3.1 Introduction

This chapter consists of descriptions on the two main methods used in this research. The first description is on the Latent Semantic Indexing which is an approach related to the reduction of matrix dimension for document representation. The second is on the fuzzy clustering algorithm (Fuzzy c-Means), that will be used to cluster the document collections.

In document categorization, the text representation plays an important role. Different representations emphasize different aspect of a document and, therefore, it is important to get representations that will generate relevant information. One of the common representation is based on the frequent terms which occurs in a given document collection. However, the dimension can be very large depending on the number of distinct terms. Moreover, a representation which is merely based on frequently occurring terms can give irrelevant results since most words can have multiple meanings. Therefore terms in a user's query which literally match terms in documents may not match with a relevant document.

3.2 Information Retrieval

Information Retrieval (IR) can be defined as an art and science for searching information in documents, searching for document themselves, searching for meta-data which describe document, or searching within databases, whether the relational stand alone databases or hypertext networked databases for text, sound, image, and data. One of the clustering functions is to help for information retrieval in document collection. The clustered document in the similar content can help user to retrieve information. The concept of information retrieval is some documents or records

containing information have been organized in order to appropriate for easy retrieval [35]. Therefore, the large data volume of document representation needs to be organized using information retrieval method without making different meaning or misinterpretation.

Information retrieval is known as a way to search the matching information in document collection as desired by a user (query). Unrelated document may be simply retrieved because the terms occur accidentally in it, and the related documents, on the other hand, may be missed because no term in the document occurs in the query. Therefore, the retrieval could be based on concept rather than on terms, by mapping first items to a "concept space" and using similarity ranking as shown in Figure 3.1 [36].

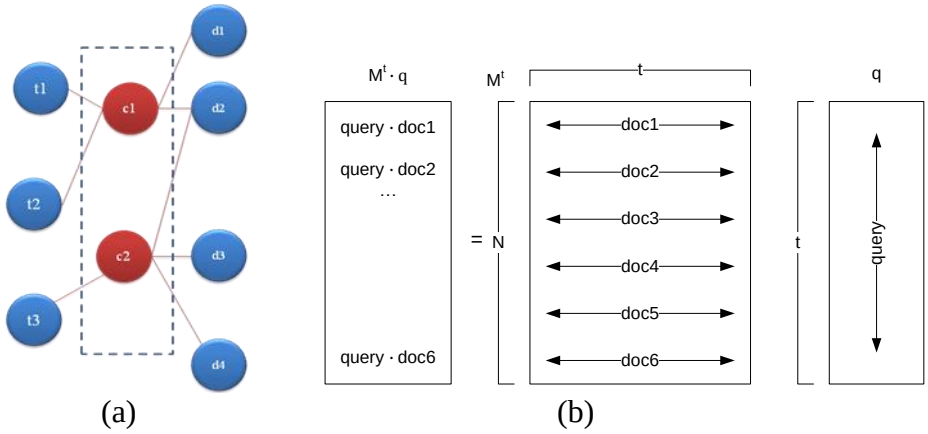


Figure 3.1 (a) Using concept for retrieval and (b) similarity ranking of document

Figure 3.1 (a) describes that there is a middle layer into two queries based on the concept (*c1* and *c2*) and document (*d1*, *d2*, *d3*, and *d4*) maps instead of directly relating documents and terms (*t1*, *t2*, *t3*) as in vector retrieval. This vector, the query *c2* of *t3* returns *d2*, *d3*, *d4* in the answer set based on the observation that they have a relationship to concept *c2*, without directly requiring the document which contains term *t3*. The middle layer is as a concept embedded in a query to retrieve the related document which is obtained using matrix computation. Thus, Figure 3.1 (b) shows how the term-document matrix *M* can be used to compute the ranking of documents with respect to a query *q*. By using similarity measurement between document and query in concept space, the document can be ranked in ascending order.

In representation document, a term-document matrix is counted from a document collection using a certain weighting scheme to be represented into vector space model. For being efficient in document representation, the concept space is applied in retrieval vector which uses a method called Latent Semantic Indexing.

3.2.1 Latent Semantic Indexing

In Information Retrieval, there is a method to get pattern in the document collection which is called a Latent Semantic Indexing (LSI). In document collection, there is a possibility of some words to have the same meaning (synonymy) and one word has many meanings (polysemy). Some words exist in documents concurrently interpreted as the same meaning. This principle uses Latent Semantic Index (LSI) to avoid any synonymy and polysemy. This method is initially applied to improve the accuracy of retrieval information. The basic concept of the LSI is to decompose the original term-document matrix of vector space model and to keep only the *k* largest singular value from singular value matrix. There are three matrices formed: two orthogonal matrices (*U* and *V*) and one diagonal matrix ($\sum$). In this matrix, $\sum$ selects the largest singular value only which keeps the corresponding columns in matrix *U* and V^T.

The selection of *k* determines the number of "essential concepts" in which its ranking will be based on. It is assumed that the concepts of small singular value of $\sum$ are considered as "noise", and thus it can be ignored. Therefore, LSI depicts how the sizes of the involved matrices are reduced, and the first *k* singular values are kept for the computation of the ranking and also how the position exists between term and document in the matrix [36]. Figure 3.2 depicts LSI implementation of data collection. The line is representative as pattern of the data collection which is plotted using dots.

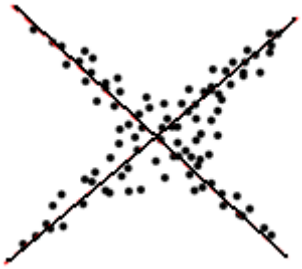


Figure 3.2 Getting pattern in data collection

The purposes of information retrieval are to rank the documents (retrieved) in order of decreasing similarity. This rank uses a query by computing which normalized inner products using cosine similarity on the term-based vectors. Principally, this method uses a mathematical approach with matrix decomposition (matrix factorization) to represent the terms in document collections and furthermore. Generally the process for LSI in information retrieval consists of the following steps [37]:

1. Applying query and text document in matrix space
2. Calculating the similarity between document and query in concept space using matrix decomposition properties
3. Sorting and ranking the relevant document by similarity

3.2.1.1 Singular Value Decomposition (SVD)

The document collections can be represented by a term-document matrix. Suppose we have *m* terms and *n* documents, we can create *m x n* term-document as matrix *A* of weighted term. However, every word (term) does not commonly appear in each document, it makes sparse matrix (zeros entries). In LSI, the matrix *A* is factored into three matrixes using Singular Value Decomposition (SVD). SVD is a method of LSI to find the patterns in the matrix and identify words and documents which are similar to each other. In this section, we outline some basic properties of singular value decomposition (SVD) used as a method of dimensionality reduction. It creates the new matrices from term (*t*) x document (*d*) matrix *A* which are matrices U, $\sum$ and V such that $A = U\sum V^T$ which can be illustrated as in Figure 3.3 [38].

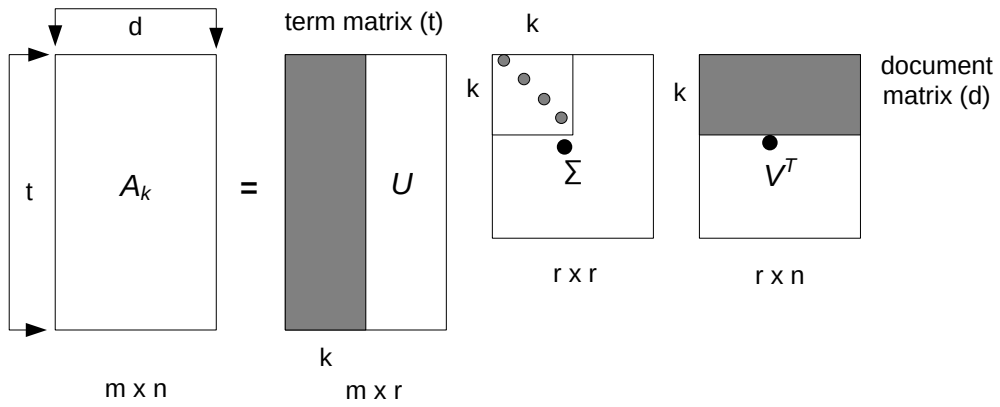


Figure 3.3 Reduced dimension representation of term-document matrix

In Figure 3.3, the SVD matrix shows the place where U and V are unitary matrices. U is an orthogonal matrix and has a unit-length column ($U^TU=I_m$). It is a matrix whose columns are the eigenvectors of the AA^T matrix and are called the left eigenvectors. V is the orthogonal matrix and has unit-length column ($V^TV=I_n$). It is a matrix whose columns are the eigenvectors of the A^TA matrix which are called the right eigenvectors. $\sum$ is a diagonal matrix of singular values, diag[$\boldsymbol{\sigma}_1$, $\boldsymbol{\sigma}_2$, …,$\boldsymbol{\sigma}_{\min\{m,n\}}$], where $\boldsymbol{\sigma}_i > 0$ for $1< i \leq r$, $\boldsymbol{\sigma}_j = 0$ for $j > r$ and r is the rank of A ($\leq$ min (m, n)). Generally, $A = U\sum V^T$ matrix has to be all of full rank. The amount of dimension reduction needs to be selected correctly by the purpose to describe the real structure in the data.

LSI constructs low rank approximation of matrix A via SVD truncation method. Supposed we have an integer k and << min (m, n). Matrix U_k can be determined to be the first k column of U, and $\boldsymbol{V_k^T}$ to be the first k rows of V^T. It is also for diagonal matrix $\sum$, and it can be defined $\sum_k$= diag[$\boldsymbol{\sigma}_1$, $\boldsymbol{\sigma}_2$, …,$\boldsymbol{\sigma}_k$] which consist of the first k largest singular values. So a new pseudo term-document matrix with reduced dimension can be defined as Equation (3.1).

$$\boldsymbol{A_k = U_k \sum_k V_k^T} \qquad (3.1)$$

Several researches about SVD truncation of LSI method have been conducted and reported satisfactorily in term of accuracy for the retrieval with other information retrieval method [39].

3.2.1.2 Principal Component Analysis (PCA)

Another way of matrix dimension reduction in LSI is the use of principal component analysis (PCA). The main objective of PCA is to get a new set of dimensions (attributes) that better captures the variability of the data. The first dimension is selected to capture as much of the variability as possible. The second dimension is orthogonal to the first dimension which captures as much of the remaining variability as possible, and so on. Hence, the strongest pattern is found in the first dimension [1].

The PCA works by finding the eigenvalues of the covariance document-term matrix and using the eigenvectors as the linear transformations to obtain the principal components. The eigenvector associated with a large eigenvalue indicates the direction in which the data has the most variance. In other words, the data is projected onto the line defined by this vector, and the result value would have the maximum variance. The significance of each principal component is determined by the relative magnitude of its eigenvalue which is as the first principal component for the highest eigenvalue. Then, the second highest is related to the second principal component, and so on.

There is correlation between PCA and SVD. Let us return to the original $n \times m$ data matrix A. The new matrix Y as an $n \times m$ matrix can be defined as in Equation (3.2) [40].

$$\boldsymbol{Y} \equiv \frac{\mathbf{1}}{\sqrt{\boldsymbol{n}}}\boldsymbol{A}^{\boldsymbol{T}} \tag{3.2}$$

where zero mean is in each column of Y. The selection of Y becomes clear by analyzing Y^TY.

$$\boldsymbol{Y}^{\boldsymbol{T}}\boldsymbol{Y} = \left(\frac{\mathbf{1}}{\sqrt{\boldsymbol{n}}}\boldsymbol{A}^{\boldsymbol{T}}\right)^{\boldsymbol{T}}\left(\frac{\mathbf{1}}{\sqrt{\boldsymbol{n}}}\boldsymbol{A}^{\boldsymbol{T}}\right)$$

$$= \frac{\mathbf{1}}{\boldsymbol{n}}\boldsymbol{A}\boldsymbol{A}^{\boldsymbol{T}}$$

$$\boldsymbol{Y}^{\boldsymbol{T}}\boldsymbol{Y} = \boldsymbol{C}_{\boldsymbol{A}} \text{ (Covariance matrix of A)}$$

The covariance matrix of A is constructed from Y^TY, and the principal components of A are eigenvectors of C_A. When the SVD of Y is calculated, the columns of matrix V consist of the eigenvectors of $Y^TY = C_A$. Therefore, the V columns are called as the principal components of A.

The steps in the construction of the principal component analysis are given in the summary below:

1. Construct data as an $n \times m$ matrix, where n is the number of measurement types (terms) and m is the number of documents
2. Subtract off the mean for each measurement type
3. Compute the SVD or the eigenvectors of the covariance.

3.2.2 Applying Latent Semantic Indexing into Information Retrieval

One of LSI's purposes is to improve the accuracy in information retrieval by getting the pattern, so that the relevant information can be easily found. Furthermore, we will describe how the LSI is applied to information retrieval using a query [40].

As an illustration assume a data set consisting of three documents (D_1, D_2, and D_3):

D_1: Shipment of gold damaged in a fire

D_2: Delivery of silver arrived in silver truck

D_3: Shipment of gold arrived in truck

The above documents can be transformed to the Term Count Model in VSM (Vector Space Model) matrix to get term weights and query weights. So, local weights are defined as word occurrences. It also uses document indexing rules as listed below:

- ignore stop words
- use text tokenized
- do not use stemming
- sort terms alphabetically

This just shows an illustration how the LSI works, even though the most modern theory and my research do not follow the above rules to get the meaningful and simplify data. The most current LSI models are not based on local weights only, but also on models that incorporate local, global, and document normalization weights. Besides that, this instance ignores stop words and terms that occur once in documents. The stemming is also not applied. There are six steps to apply the LSI on the above three documents of the rank for the query (q): *'gold silver truck'.*

The first step is to keep count term weights and build the term-document matrix A and query matrix q. The number of this illustration states the frequency of each term, although this research uses TF-IDF weighted term.

Terms	d_1	d_2	d_3	Q
A	1	1	1	0
arrived	0	1	1	0
damaged	1	0	0	0
delivery	0	1	0	0
fire	1	0	0	0
gold	1	0	1	1
in	1	1	1	0
of	1	1	1	0
shipment	1	0	1	0
silver	0	2	0	1
truck	0	1	1	1

Then, the decomposition of matrix *A* is applied. It will be obtained three factorial matrices from decomposition (SVD) of matrix *A* such as *U*, ∑ and *V* matrices, where matrix $A=U\sum V^T$.

$$\begin{bmatrix} 1 & 1 & 1 \\ 0 & 1 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 0 & 1 \\ 0 & 2 & 0 \\ 0 & 1 & 1 \end{bmatrix} = \begin{bmatrix} U1 & U2 \end{bmatrix} \times \begin{bmatrix} \sigma_1 & \Phi \\ \Phi & \sigma_2 \end{bmatrix} \times \begin{bmatrix} V1 \\ V2 \end{bmatrix}$$

(A) (U) (∑) (V)

By refer to the properties of decomposition matrix, it can be generated as follows:

- $U = A^T A_x$; $V = AA^T{}_x$; $\boldsymbol{\sigma} = \sqrt{\boldsymbol{\lambda}}$

  where λ is eigenvalue, X is eigen vector and $\boldsymbol{\sigma}$ is singular value
- $A\,\lambda = X\,\lambda$
- $det(A - \lambda I) = 0$

Then, ∑, U and V are directly obtained as the next step.

U =

$$\begin{bmatrix} -0.4201 & 0.0748 & -0.0460 \\ -0.2995 & -0.2001 & 0.4078 \\ -0.1206 & 0.2749 & -0.4538 \\ -0.1206 & -0.3046 & -0.2006 \\ -0.1576 & 0.2749 & -0.4538 \\ -0.2626 & 0.3794 & 0.1547 \\ -0.4201 & 0.0748 & -0.0460 \\ -0.4201 & 0.0748 & -0.0460 \\ -0.2626 & 0.3794 & 0.1547 \\ -0.3151 & -0.6093 & -0.4013 \\ -0.2995 & -0.2001 & 0.4078 \end{bmatrix}$$

$$\sum = \begin{bmatrix} 4.0989 & 0.0000 & 0.0000 \\ 0.0000 & 2.3616 & 0.0000 \\ 0.0000 & 0.0000 & 1.2737 \end{bmatrix}$$

V =

$$\begin{bmatrix} -0.4945 & 0.6492 & -0.5780 \\ -0.6458 & -0.7194 & -0.2556 \\ -0.5817 & 0.2469 & 0.7750 \end{bmatrix}$$

V^T =

$$\begin{bmatrix} -0.4945 & -0.6458 & -0.5817 \\ 0.6492 & -0.7194 & 0.2469 \\ -0.5780 & -0.2556 & 0.7750 \end{bmatrix}$$

The third step is to apply a rank with three approximations by keeping the first columns of *U* and *V* and the first columns and rows of $\sum$. Because of the small document size, a selected rank of 2 (*k*=2); select an optimal rank is still in issue within the research community. The rank of those term-document matrix is the total element in diagonal matrix diagonal and it is minimum of total document (*n*) and total different term (*m*) notated as $k \leq (m, n)$.

$U \approx U_k =$

$$\begin{bmatrix} -0.4201 & 0.0748 \\ -0.2995 & -0.2001 \\ -0.1206 & 0.2749 \\ -0.1206 & -0.3046 \\ -0.1576 & 0.2749 \\ -0.2626 & 0.3794 \\ -0.4201 & 0.0748 \\ -0.4201 & 0.0748 \\ -0.2626 & 0.3794 \\ -0.3151 & -0.6093 \\ -0.2995 & -0.2001 \end{bmatrix}$$

$\sum \approx \sum_k =$

$$\begin{bmatrix} 4.0989 & 0.0000 & 0.0000 \\ 0.0000 & 2.3616 & 0.0000 \\ 0.0000 & 0.0000 & 1.2737 \end{bmatrix}$$

$V \approx V_k =$

$$\begin{bmatrix} -0.4945 & 0.6492 \\ -0.6458 & -0.7194 \\ -0.5817 & 0.2469 \end{bmatrix}$$

$V^T \approx V^T_k =$

$$\begin{bmatrix} -0.4945 & -0.6458 & -0.5817 \\ 0.6492 & -0.7194 & 0.2469 \end{bmatrix}$$

The next step is to find the new document vector coordinates in a reduced 2-dimensional space. Rows of *V* hold eigenvector values. These are the coordinates of individual document vectors, and are obtained:

d1 (-0.4945, -0.6458)

d2 (-0.6458, -07194)

d3 (-0.5817, 0.2469)

The fifth step is to find the new query vector coordinates in the reduced 2-dimensional space $q = q^T U_k \sum_k^{-1}$. There is a new coordinate of query vector in two dimensions. This matrix is different from the original query matrix *q* as shown in the first step.

$$q = [0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 1 \quad 0 \quad 0 \quad 0 \quad 1 \quad 1] \begin{bmatrix} -0.4201 & 0.0748 \\ -0.2995 & -0.2001 \\ -0.1206 & 0.2749 \\ -0.1576 & -0.3046 \\ -0.1206 & 0.2749 \\ -0.2626 & 0.3794 \\ -0.4201 & 0.0748 \\ -0.4201 & 0.0748 \\ -0.2626 & 0.3794 \\ -0.3151 & -0.6093 \\ -0.2995 & -0.2001 \end{bmatrix} \begin{bmatrix} \frac{1}{4.0989} & 0.0000 \\ 0.0000 & \frac{1}{2.3616} \end{bmatrix}$$

$$q = [-0.2140 \quad -01821]$$

Finally, we rank documents in decreasing order of query-document using cosine similarities. The general formula of cosine similarities is:

$$sim(query, doc) = \frac{query \cdot doc}{|query| \cdot |doc|}$$

$$sim(q, d1) = \frac{(-0.2140)(-0.4945) + (-0.1821)(0.6492)}{\sqrt{(-0.2140)^2 + (-0.1821)^2}\ \sqrt{(-0.4945)^2 + (0.6492)^2}} = -0.0541$$

$$sim(q, d2) = \frac{(-0.2140)(-0.6458) + (-0.1821)(-0.7194)}{\sqrt{(-0.2140)^2 + (-0.1821)^2}\ \sqrt{(-0.6458)^2 + (-0.7194)^2}} = 0.9910$$

$$sim(q, d3) = \frac{(-0.2140)(-0.5817) + (-0.1821)(0.2469)}{\sqrt{(-0.2140)^2 + (-0.1821)^2}\ \sqrt{(-0.5817)^2 + (0.2469)^2}} = 0.4478$$

After the similarity measurement is obtained, the documents can be ranked in descending order as shown by: *d2 > d3 > d1*. The document *d2* scores higher than *d3* and *d1*. The document which has high similarity, means having contiguous vectors. It is closer to the query vector than the other vectors as shown in Figure 3.4.

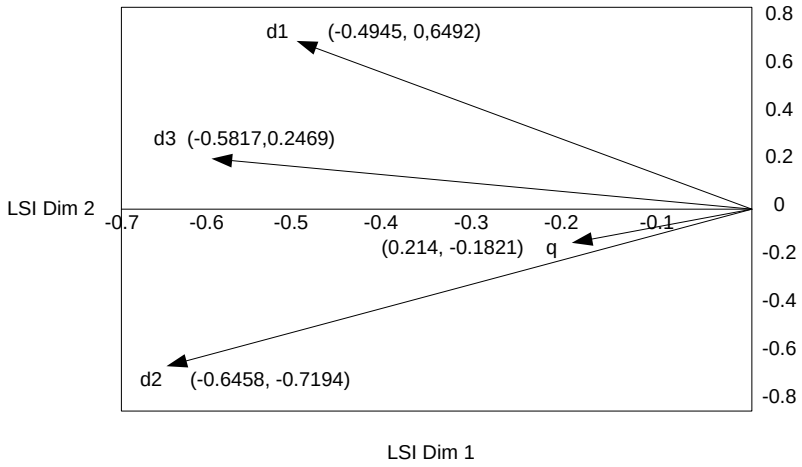


Figure 3.4 Document and query vector in similarity concept

3.3 Fuzzy Clustering

Fuzzy clustering is a grouping method based on the similarity of the characteristic of data using membership degree. It is different from crisp (hard) clustering which each datum is assigned to exactly one cluster. The fuzzy clustering is represented by grades of membership of every pattern to the classes established by evaluating in the [0, 1] intervals. In other words, the fundamental fuzzy clustering is to consider not only the belonging class (status) to the clusters, but also to what degree do the objects belong to the clusters [38]. The membership function on X represents fuzzy subsets of X which represents a fuzzy set $\tilde{A}$ that is usually denoted by $\mu_A(x)$ as shown in Figure 3.5 [41].

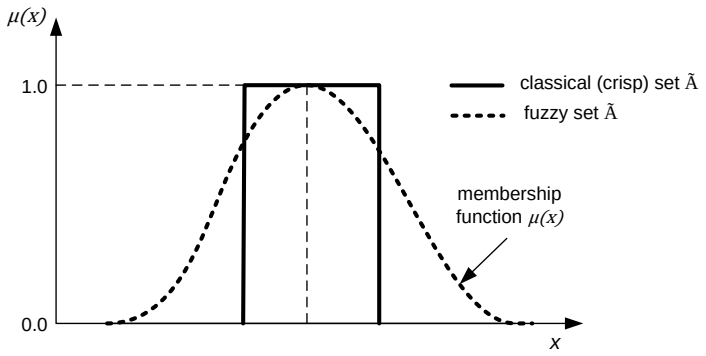


Figure 3.5 Membership function of fuzzy set

In mathematical notation, the sample data can be written as $\boldsymbol{X} \in \mathfrak{R}^{\boldsymbol{pxn}}$, where *n* is the number of collected data and *p* denotes the number of feature attributes, variables or measured quantities. One datum is denoted as the vector $\boldsymbol{X} \in \mathfrak{R}^{\boldsymbol{pxn}}$ with $\boldsymbol{k} \in \{\boldsymbol{1}, \dots, \boldsymbol{n}\}$ and the data as set of feature vectors: $\boldsymbol{X} = \{\boldsymbol{x_1}, \boldsymbol{x_2}, \dots, \boldsymbol{x_n}\}$. As the goal of cluster is to get partition of the sample data into subgroups denoted by $C = \{v_1, \dots, v_c\}$, where *c* is the number of clusters or subgroups. The covariance matrix is a parameter which influences the dissimilarity or distance measure $d^2(v_i, x_k)$. Furthermore, each data vector $\boldsymbol{x_k} \in \boldsymbol{X}$ is assigned to each cluster vector $\boldsymbol{v_i} \in \boldsymbol{C}$. The membership degree is a grade of data and denoted by $\boldsymbol{\mu_{ik}}$. The matrix of all membership degree is showed in Equation (3.3).

$$\boldsymbol{U} = \begin{bmatrix} \boldsymbol{\mu_{1,1}} & \cdots & \boldsymbol{\mu_{1,c}} \\ \vdots & \ddots & \vdots \\ \boldsymbol{\mu_{N,1}} & \cdots & \boldsymbol{\mu_{N,c}} \end{bmatrix} \tag{3.3}$$

Hard clustering admits only crisp membership degrees $(\boldsymbol{\mu_{ik}})$ either 0 or 1, $\boldsymbol{\mu_{ik}} \in \{\boldsymbol{0}, \boldsymbol{1}\}$, however in fuzzy clustering the membership degree lies b between 0 and 1. In other words, the membership degree of fuzzy clustering is noted as $\boldsymbol{\mu_{ik}} \in [\boldsymbol{0} \dots \boldsymbol{1}]$. The clustering task is to determine a partition using similarity or distance concept, where similar data is grouped in one cluster if the degree of membership of the data object is close to one ($\boldsymbol{\mu_{ik}} \rightarrow \boldsymbol{1}$), for small dissimilarity $d^2(v_i, x_k)$. In contrast, large dissimilarity means the data is partitioned in different groups. The fuzzifier or fuzziness index denoted by $\boldsymbol{m} \in \boldsymbol{\mathfrak{R}_{>1}}$, is used as an exponent of the membership degree. For $\boldsymbol{m} \rightarrow \boldsymbol{1}$, the membership degrees are like hard clustering, either 0 or 1. However, if the fuzziness is $\boldsymbol{m} \rightarrow \infty$, then the membership is $\frac{\boldsymbol{1}}{\boldsymbol{c}}$. In this case, datum is totally split among the clusters and assigned it with the same degree to each cluster.

Thus, a value of 2 for *m* is often chosen [42], but *m* is also chosen to close 1 to return back hard from fuzzy clustering method [21, 25]. As we know that the task of clustering is to determine a partition where similar data is grouped in one cluster.

In general, the objective function of fuzzy clustering is the summing up of dissimilarity weighted (d) by membership degree (μ) as shown in Equation (3.4) [43]:

$$\boldsymbol{J(X;U,V) = \sum_{i=1}^{c}\sum_{k=1}^{N}(\mu_{ik})^{m}\, d^{2}(v_i, x_k)} \qquad (3.4)$$

To obtain in minimizing the objective function, it is given constraint as declared in Equation (3.5) which guarantees that no cluster is empty (without data) and another constraint in Equation (3.6) ensures that each datum has the same total influence of membership degrees (μ) equal to 1 .

$$\boldsymbol{\sum_{k=1}^{n}\mu_{ik} > 0 \; for\; all\; i \in \{1, \dots, c\}} \qquad (3.5)$$

$$\boldsymbol{\sum_{i=1}^{c}\mu_{ik} = 1 \; for\; all\; k \in \{1, \dots, n\}} \qquad (3.6)$$

The last constraint in Equation (3.6) is usually known as probabilistic fuzzy clustering because of the membership degrees from a datum is like the probability of it being a member of clusters. The property of a probabilistic clustering is partitioning, which 'distributes' the weight of a datum to the different clusters due to the restriction (constraint).

However, the objective function *J* is not able to minimize directly. Therefore, an iterative algorithm needs to be used, which the membership degrees and the cluster parameters are alternately optimized [42, 43]. Initially, the membership degrees are optimized for fixed cluster parameters, and then the cluster parameters are optimized for fixed membership degrees. The main advantage of this is that in each of the two steps the optimum can be computed directly. By iterating the two steps the joint optimum is approached (although, it cannot be guaranteed that the global optimum will be reached, the algorithm may be immovable in a local minimum of the objective function *J*). The update formulas are derived by simply setting the derivative of the objective function *J*. The parameters used to optimize so that the derivative of objective function equal to zero (necessary condition for a minimum). The chosen

distance measure is as an independent, which the memberships are not restricted by small distances to a different cluster but by growing distance to all cluster, thus it can be obtained the following update formula for the membership degrees [42]. The membership represents the relative inverse squared distances of data point to the different cluster centers, which is a very intuitive result (using iterative procedure). The membership values of the projected data can be constructed based on the classical formula of the calculation of the membership values as denoted in Equation (3.7):

$$\mu_{ik} = \frac{1}{\sum_{j=1}^{c} \left(\frac{d^2(V_i, X_k)}{d^2(V_j, X_k)} \right)^{\frac{1}{m-1}}} \tag{3.7}$$

This is general fuzzy clustering concept. However, getting different fuzzy algorithms relies on the definitions of suitable distance measures. Choosing a certain dissimilarity measurement defines the structure, which is searched for in the sample data. In addition, different distance measures are able to describe varying forms or shapes of clusters. In the next section, we describe the most common fuzzy clustering algorithm, Fuzzy c-Means that is used for clustering in this research.

3.3.1 Fuzzy c-Means Clustering

There are various fuzzy clustering algorithms. One of the simplest fuzzy clustering techniques is the Fuzzy c-Means algorithm (FCM) by Dunn (1973) which is called as the birth of fuzzy method. The clusters are assumed to be of approximately the same size. Each cluster is represented by its center. This algorithm is then developed by Bezdek who introduces the fuzzifier *m* [43]. The FCM is applied in this research in order to cluster document collections due to its ability to produce reasonable partitioning of the original data in many cases and its quick processing time compared to other approaches [43]. For example, [44] has shown that FCM is, on average, perhaps an order of magnitude faster than the maximum likelihood approach for estimation of the parameters of a mixture of two univariate normal distributions. Besides, FCM always converges to a minimum or a stationary point from any initialization; it may either be a local or global minimum of objective function [45].

Furthermore, it is able to produce the best clusters by identifying the cluster centroid and their corresponding degree of membership until the threshold is minimized [46].

Fuzzy c-Means algorithm recognizes spherical clouds of points in a *p*-dimensional space. Supposed the clusters are approximately the same size. It is represented by its center in each cluster. This representation of a cluster is also called a prototype v_i as in Eq. 3.9, since it is often regarded as a representative of all data assigned to the cluster. Thus, as a measure for the distance in this algorithm (FCM), it is used the Euclidean distance between a datum and a prototype is used Equation (3.8) [43].

$$\boldsymbol{D_{ikA}^2 = \|x_k - v_i\|_A^2 = (x_k - v_i)^T A (x_k - v_i)} \tag{3.8}$$

$$\boldsymbol{v_i = \frac{\sum_{k=1}^{N} \mu_{ik}^m x_k}{\sum_{k=1}^{N} \mu_{i,k}^m}, \qquad 1 \le i \le c} \tag{3.9}$$

where x_k is an object of data set, A is the covariance matrix, μ_{ik} is the membership for the k object of a data set, c is the number of the clusters.

Algorithm

Given the data set X, select the number of clusters $1 < c < N$, the weighting exponent $m > 1$, the termination tolerance $\boldsymbol{\varepsilon} > 0$ and the norm-inducing matrix A. Initializing the partition matrix randomly, such that $U^{(0)}$, repeat for $l = 1, 2, \ldots$where l is an iteration.

Step 1: Count the cluster prototype (means) v_i :

$$\boldsymbol{v_i^{(l)} = \frac{\sum_{k=1}^{N} (\mu_{ik}^{(l-1)})^m x_k}{\sum_{k=1}^{N} (\mu_{i,k}^{(l-1)})^m}, \qquad 1 \le i \le c}$$

where ,

μ_{ik} = membership value of the compound x_k in cluster i

Step 2: Count distances $\boldsymbol{D_{ik}^2}$

$$\boldsymbol{D_{ikA}^2 = (x_k - v_i)^T A (x_k - v_i), \qquad 1 \le i \le c, \qquad 1 \le k \le N}$$

where,

$\boldsymbol{D_{ik}^2}$ the distance between feature vector x_k and prototype v_i

A any positive definite matrix which in the case of Euclidean distance is the identity matrix

Step 3: Update the membership (partition matrix):

$$\boldsymbol{\mu}_{i,k}^{(l)} = \frac{\mathbf{1}}{\sum_{j=1}^{c}(\boldsymbol{D}_{ikA}/\boldsymbol{D}_{jkA})^{2/(m-1)}}$$

Until $\left\|\boldsymbol{U}^{(l)} - \boldsymbol{U}^{(l-1)}\right\| < \boldsymbol{\varepsilon}$ (threshold)

where, *U is the* partition matrix for the evaluated data set. The prototypes are the norm of the membership matrices (U^{old}) and (U^{new}) is in iteration until smaller than a given bound ε.

3.3.2 Similarity Measurement in Fuzzy c-Means

In clustering system, the similarity or dissimilarity is a way to measure whether one document can be grouped to the others. This is a measurement from cluster center which has the significant words or terms. The objective function of Fuzzy c-Means is summing up of dissimilarity by membership degree of each document.

The dissimilarity of data object (document) is shown by the distance between document as cluster center and the others. One of the dissimilarity measurements is Euclidean distance as stated in Equation (3.8). Suppose we have distance *d*, between two points, *x* and *y*, in one-, two-, three-, or higher-dimensional space, then the formulation of distance can be restated as below:

$$\boldsymbol{d}(\boldsymbol{x},\boldsymbol{y}) = \sqrt{\sum_{\boldsymbol{k}=\mathbf{1}}^{\boldsymbol{n}}(\boldsymbol{X}_{\boldsymbol{k}} - \boldsymbol{Y}_{\boldsymbol{k}})^{\mathbf{2}}}$$

where, *n* is the number of dimensions and X_k and Y_k are respectively, the k^{th} attributes (components) of *x* and *y*. The measure of Euclidean distance is generalized by the Minkowski distance matrix as in the Equation (3.10).

$$\boldsymbol{d}_{\boldsymbol{p}}(\boldsymbol{x}_{i,k} - \boldsymbol{x}_{j,k}) = \left(\sum_{\boldsymbol{K}=\mathbf{1}}^{\boldsymbol{d}}\left|\boldsymbol{x}_{i,k} - \boldsymbol{x}_{j,k}\right|^{\boldsymbol{p}}\right)^{\frac{\mathbf{1}}{\boldsymbol{p}}} \qquad (3.10)$$

where, *d* is the number (dimensionality) of the data. The Euclidean distance is *p*=2 (L_2 norm), while Manhattan metric has *p*=1 (L_1 norm)and $p = \infty$, it is known as Supremum distance (L_{max} or L_∞ norm) [1].

In contrast, the similarity between data object (document) is known as the small distance in one cluster. Documents are often represented as vectors, where each attribute represents the frequency in which a particular term (word) occurs in the document. A measure of similarity for document clustering is the cosine of the angle between two vectors as in this Equation (3.11) [1].

$$\cos(\boldsymbol{d_i}, \boldsymbol{d_j}) = \frac{\boldsymbol{d_i^t} \cdot \boldsymbol{d_j}}{\|\boldsymbol{d_i}\| \cdot \|\boldsymbol{d_j}\|} \qquad (3.11)$$

For instance, calculation of the following two data objects, which might represent document vectors, d_i and d_j.

d_i = (3,2,0,5,0,0,0,2,0,0) ; d_j = (1,0,0,0,0,0,0,1,0,2)

$d_i \cdot d_j$ = 3*1+2*0+0*0+5*0+0*0+0*0+0*0+2*1+0*0+0*2 = 5

$\|\boldsymbol{d_i}\|$

$= \sqrt{\mathbf{3*3+2*2+0*0+5*5+0*0+0*0+0*0+2*2+0*0+0*0}} = \mathbf{6.48}$

$\|\boldsymbol{d_j}\|$

$= \sqrt{\mathbf{1*1+0*0+0*0+0*0+0*0+0*0+0*0+1*1+0*0+0*0+2*2}} = \mathbf{2.24}$

Cos (d_i,d_j)=0.31

Thus, if the cosine similarity of the objects is 1, the angle is between d_i and d_j is 0^0, also d_i and d_j are the same except for magnitude (length). In contrast, if cosine similarity is 0, then the angle is between *x* and *y* is 90^0, and they do not share any terms (words). Furthermore, the cosine similarity can be shown at Figure 3.6.

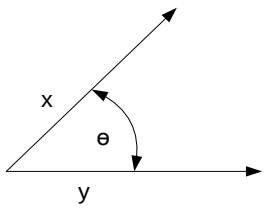


Figure 3.6 Geometric illustration of the Cosine measure

And to know the comparison between Euclidean and Cosine properties can be illustrated in Figure 3.7.

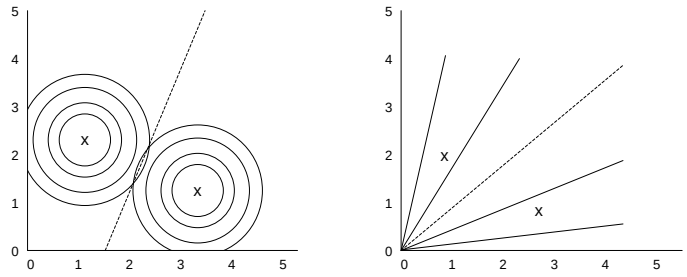


Figure 3.7 Properties of (a) Euclidean-based, (b) Cosine similarity illustrated in two dimension.

The Euclidean distance $\boldsymbol{D = \|x_k - v_i\|^2}$ is a dissimilarity measurement between a data object x and a cluster center *v*. There are also other types of measurements such as Manhattan distance (city-block). Another class contains similarity measures instead of dissimilarity: a similarity measure between arbitrary pair $\boldsymbol{x, x' \in X = \{x_1, \dots, x_N\}}$ is denoted by *S{x, x'}* which takes a real value, which is symmetric with respect to the two arguments [47]:

$$\boldsymbol{S(x, x') = S(x', x), \quad \forall x, x' \in X} \tag{3.12}$$

In contrast to a measure of dissimilarity, a large value of $\boldsymbol{S(x, x')}$ means *x* and *x'* are near, while a small value of $\boldsymbol{S(x, x')}$ is distant. Therefore, in particular, we can assume

$$\boldsymbol{S(x, x) = \max_{x' \in X} S(x', x).}$$

Assume the inner product of Euclidean space $\boldsymbol{R^p}$ be

$$\boldsymbol{\langle x, y \rangle = x^T y = y^T x = \sum_{j=1}^{p} x^j y^j}$$

then cosine correlation is defined by:

$$\boldsymbol{S_{cos}(x, y) = \frac{\langle x, y \rangle}{\|x\| \|y\|}} \tag{3.13}$$

So, the dissimilarity can be defined from $\boldsymbol{S_{cos}(x, y)}$ as:

$$\boldsymbol{D(x, v) = 1 - S(x, v)} \tag{3.14}$$

Then, this formula is applied to the objective function of Fuzzy c-Means algorithm:

$$J'(U,V) = \sum_{i=1}^{c}\sum_{k=1}^{N}(\mu_{ki})^m D(x_k, v_i)$$

$$J'(U,V) = \sum_{i=1}^{c}\sum_{k=1}^{N}(\mu_{ki})^m (1 - S(x_k, v_i)) \quad (3.15)$$

We, thus, employ *J'(U, V)* in FCM algorithm and alternative minimization should be done as usual.

$$U_{ki} = \left[\sum_{j=1}^{c}\left(\frac{D(x_k, \overline{v}_i)}{D(x_k, \overline{v}_j)}\right)^{\frac{1}{m-1}}\right]^{-1} = \left[\sum_{j=1}^{c}\left(\frac{1 - S(x_k, v_i)}{1 - S(x_k, v_j)}\right)^{\frac{1}{m-1}}\right]^{-1}$$

While $\overline{V}$ in FCM is

$$\overline{v}_i = \frac{\sum_{k=1}^{N}(\overline{\mu}_{ki})\frac{x_k}{\|x_k\|}}{\left\|\sum_{k=1}^{N}(\overline{\mu}_{ki})\frac{x_k}{\|x_k\|}\right\|} \quad (3.16)$$

3.3.3 Fuzzy Prediction Methods

In fuzzy clustering algorithm, the result is the membership degree of each datum which has been clustered. To evaluate the cluster result, Holiday et al. (2003) classifies that there are four approaches of fuzzy prediction method (PM) and the details are seen as below:

1. PM1 is an approach that can be used to 'defuzzify' the fuzzy partition after clustering has taken place. Each document is assigned to a cluster which has the highest membership function (the 'home' cluster) and assigned $\mu=1$ for that cluster and $\mu=0$ for all other clusters.

2. PM2 is an approach that predict to fuzzy partition using these steps:
    a. Assign each document its home cluster as above
    b. Apply a threshold to the data, μ_1, below which membership to cluster is ignored
    c. Address cluster of document *j* into cluster *i*

d. Calculate the maximum of property value for cluster *i* by summing each qualifying membership function ($\mu > \mu_1$) for the property value of that document (excluding document *j*)
e. Predict property for *j* from the maximum value in clusters *i*.

3. PM3 is an approach of prediction to the result of fuzzy partition using some steps as follows:
    a. The membership functions are included in the prediction stage
    b. Take each cluster in turn
    c. Calculate the maximum value for each cluster by summing up membership function for each document.
    d. Sum up the property values of document *j* for that cluster to obtain an overall property value for *j*
    e. The membership function of each document is being considered and thus each document contributes more or less to the prediction depending on the degree on which it is a member of specific cluster.

4. Then, PM4 is also an approach of prediction in fuzzy partition result. The step details are below:
    a. This approach is an extension of PM3
    b. Make prediction from selecting a minimal membership or the minimum membership function
    c. Put membership functions in descending order for document *j*. Then, count the summation of these membership is reached, these are the clusters that will be used for prediction
    d. Sum up the property values of document *j* for that cluster to obtain an overall property value for *j*

The membership function of each document is being considered and thus each document contributes more or less to the prediction depending on the degree to which it is a member of specific cluster.

3.4 Summary

This chapter has explained the background theory used for the proposed method. There are two methods that can be applied to cluster documents which adopted from information retrieval. First, Latent Semantic Indexing is a method used to improve document representation based on concept space in vector domain. The second method is a basic concept of fuzzy clustering algorithm itself and its extension, especially for the majority of this algorithm. The cosine similarity which uses information retrieval in ranking similar document is applied to replace Euclidean distance formula in objective function of fuzzy clustering algorithm. Besides the mentioned method, a fuzzy prediction method is described to assign it to the respective cluster number.

CHAPTER 4

METHODOLOGY

4.1 Introduction

This chapter describes the methodology of document clustering into their respective clusters. The first part of this chapter discusses the *preprocessing* stage aimed at removing the common words and *stemming* it to obtain the local and global threshold for the frequent words. Subsequently, the documents are represented into Vector Space Model. After that, Latent Semantic Indexing (LSI) is applied to document representation to map document collections in concept space by identifying their trends or their patterns in a multidimensional data set. Then, fuzzy clustering algorithm itself is implemented to cluster the documents.

The proposed framework contains additional step 3 and modified clustering algorithm in step 4 from original concept of document clustering (see Section 2.5). The details proposed method are illustrated as the five stages below:

1. Preprocessing of document collections
2. Transforming document to weighting term matrix representation (Vector Space Model) using TF-IDF formula
3. Finding the patterns in the matrix representation using LSI approach in order to map the document collections into concept space.
4. Applying Fuzzy c-Means clustering algorithm
5. Assignment of documents to the respective cluster number

The third step is proposed to reduce the size of term-document matrix representation. It is used for mapping the document into concept space in vector domain before applying the clustering algorithm. In step 4, Fuzzy c-Means algorithm is employed as the fuzzy clustering method. Fuzzy c-Means clustering is a prominent fuzzy clustering algorithm. To improve the result, there is a modification of formula in its objective function as shown in highlighted area of Figure.4.1.

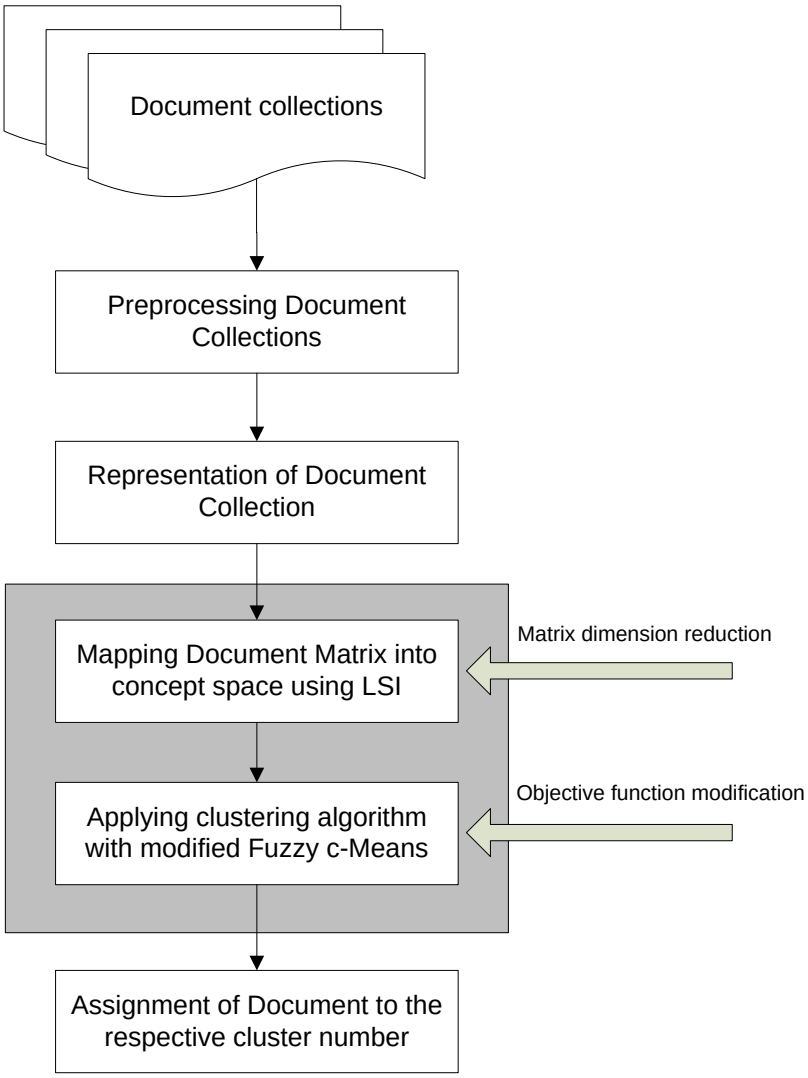


Figure 4.1 Framework of the methodology for document clustering

4.2 Preprocessing Step for Document Clustering

In real-world, data is often incomplete, noisy, and not consistent. The volume of data is also very huge with various meanings. Data *preprocessing* is an important issue before clustering by reducing the meaningless information. The data reduction techniques is a way to obtain a reduced representation of data set that is much smaller in volume, while maintaining closely integrity of the original data. In document clustering, this technique is applied to remove all the meaningless stop words; such as articles, preposition and auxiliaries verbs in the stored table so that each document is transformed into a structure that will be used by the similarity function *f()*. In short, the preprocessing of document clustering includes these following steps; Case Folding, Parsing, Stemming, Removing stop word, Giving Threshold (minimum and

maximum global frequency; minimum and maximum local frequency) of weighting term-document.

First step of *preprocessing* document is *case folding.* This step is to change all words (terms) in document collections to small case letters in order to recognize the morphology of words without any difference between small case and capital letters.

Besides that, by *parsing*, the document will be split into each term or word in order to recognize whether the term is used in the following process or to be ignored, such as numbering, white space, etc.

Another step of preprocessing is *stemming*. This step seeks to get the root word by removing their suffix. For instance, the words "presentation", "presented", "presenting" could all be decreased to a common representation of "present". This is a widely used morphological technique procedure in text processing for IR based on the assumption that posting a query with the term *presenting* implies an interest in documents containing the words *presentation* and *presented*. By using the stemming, the searcher does not need to put the correct truncation point of search keys. It reduces the total number of distinct index entries. According to Hull in 1996, the stemming is very useful to compose short query. With short queries and short documents, a derivational stemmer is the most useful so that it obtains more relevant document [48]. However, *stemming* can increase more ambiguity of search key and greedy stemming might be a counter-productive: with long queries and documents, the relevant material is identified by conservative stemming. Stemmers have been more successful in than English text retrieval such as in Slovenian, French, Modern Greek, Arabic and Swedish [49]. Lycos and Google are examples of search engines using stemming algorithms.

In stemming step, porter stemmer algorithm is used. The Porter Stemmer is a famous stemming method which is known as a conflation stemmer and introduced by Martin Porter at the University of Cambridge in 1980. This algorithm is more robust compared to other common stemming algorithms such as Lovin's and S-Stemmer. It has higher accuracy and better results compared to the Lovin's stemmer. When the performance of several stemmers on query terms is compared, and measured by their ability to find correct conflations and to find nothing but correct conflations, it must

be weighted by in-collection frequency of the terms involved. The Porter Stemmer achieved 97% accuracy at 90% coverage of potential conflations. It strongly suggests that stemming is an effective approach in solving conflations. The conflation is found at most in other than Porter algorithm [50].

Porter's algorithm was developed for English-language texts stemming but the importance of information retrieval increase in the 1990s led to a propagation of interest in the development of conflation techniques that would enhance the searching of texts written in other languages. As of now, Porter algorithm has become a standard for stemming English-language text, hence it provides a natural model to process other languages. Some of these new algorithms uses a very restricted suffix dictionary from the original algorithm [51]. Furthermore, the Porter's algorithm is important for two reasons. First, it provides a simple approach to conflation that work well in practice and is applicable to a range of different languages. The second reason is that it has spurred interest in stemming as a topic for research in its own right, rather than a merely low-level component of an information retrieval system. This algorithm is developed so that its descendants continue to be employed in a range of applications that stretch far beyond its original intended use [52].

The basic concept of the stemmer is the idea that the suffixes in the English language (approximately 1200) are mostly composed of a combination of smaller and simpler suffixes. This stemmer step is a linear. In particular, it has five stages in applying rules in each step. If a suffix rule matched to a word, in each step, then the conditions embedded to that rule are tested to the resulting stem. And when omitting that suffix, it uses in the way determined by the rule. For example, a condition with the number of vowel characters, which are followed a consonant character in the stem, then it must be greater than one for the rule to be applied. In Figure 4.2, each rule will be fired if once the condition is passed and admitted, and the suffix is omitted and control shifts to the next step. If the rule is not admitted then the next rule in the step is checked, that either a rule from that step fires and control passes to the next step or there are no more rules in that step from control moves to the next step. This process occurs on the steps, the consequential stem being returned by the Stemmer after the control has been passed step five. In short, the description for step of data preprocessing steps by using Porter's algorithm is as follows: lexical analysis,

stop word elimination that is the removal of very frequent words such as articles, prepositions, conjunctions, etc which contain little information (meaningless) of processed document. Thus, the stemming, a replacement of all variants of a word with a single common stem or another word, get the root word [53]. The illustration below shows that each step consists of set of rules in the form <condition> <suffix> → <new suffix>. In given a rule ($m > 0$) EED → EE means "if the word has at least one vowel and consonant plus EED ending, change the ending to EE". So "agreed" becomes "agree" while "feed", it will remain unchanged.

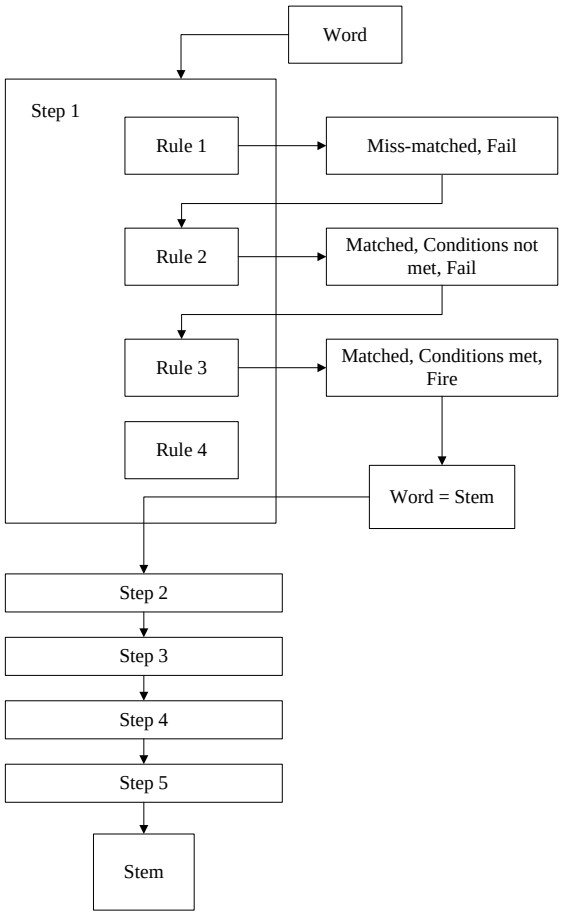


Figure 4.2 Stemming process using Porter's algorithm

Another step in preprocessing document is the stop word removal. It has been known since the earliest days of information retrieval (Luhn, 1957) that many of the most frequently occurring words in English such as "the", "of", "and", "to", etc are meaningless words as index terms. A search using one of these terms could possibly lead to the retrieval of almost every item in a database in spite of its relevance, so their intolerance value is low [54]. Furthermore, these words compose a large fraction of the text of most documents: the ten most frequently occurring words in English

typically account for 20 up to 30 percent of the tokens in a document. A list of words selected out during automatic indexes because they create poor index terms is called *a stop list* or *a negative dictionary*. The list of stop word is in Appendix A.

The last step in *preprocessing* is giving threshold of frequent term in document collection. To keep similarity of documents, we give threshold minimum and maximum frequencies of terms inside a document (local frequency) and all documents (global frequency). For example, in global frequency, we can define that the documents are in one cluster if they have minimum three frequent terms (in the same term). And they have also limitation of maximum frequent term. It aims is to remove unimportant term.

4.3 Representation of Document Collection

A document is composed of many concurrent terms so that it creates the certain pattern and meaning. In order to make easy to handle the documents and reduce their complexity, the document have to be transformed from the full text to a document vector which describes the contents of the document. To represent text document, we use Vector Space Model (VSM) where its document content is formalized as a dot in the multi-dimensional space and represented by a vector d, such as $d = \{ w_1, w_2, \ldots w_n \}$, where w_i (i = 1, 2, …, n). It is a term weight in one document. The term weight value denotes the significance of this term in a document.

The procedure of vector space model can be divided into three phases. The first phase is the document indexing, in which the non significant words are removed from the document vector, so the document which has full meaning will be presented [54]. After that, the weighting of indexing term is carried out in order to enhance the retrieval of relevant document to the user. The last phase is to sort the document with respect to a similarity measurement.

The most common representation technique of document collection in information retrieval is the vector space model by TF-IDF weighting. In vector representation, each dimension denotes the presence or absence of a certain word in that document. The term weighting of the *vector space model* has entirely used single term statistics. It is a well-known approach for computing word weights. This method assigns the

weight to each word in a document in proportion to the number of occurrences of the word in the document, and in inverse proportion to the number of documents in the collection in which the word occurs at least once.[55] This statement can be seen in the formulation as seen in Equation (4.1).

Tf - idf = term-frequency (term, document) · idf(t)

$$\boldsymbol{idf(t) = log\left(\frac{|D|}{document - frequency(t)}\right)}$$

Then, the term weight (TF-IDF) is

$$\boldsymbol{w_i = tf_i * log\left(\frac{D}{df_i}\right)} \quad (4.1)$$

where,

- tf_i = frequency of term (term counts) or how many times a term i occurs in a document. This accounts for local information
- df_i = total of documents (frequency of document) which containing term i
- D = total of documents in a database
- $log(D/df_i)$ = inverse document frequency (idf_i). This accounts for global information

The term-document matrix is counted in each document and whole document, after *preprocessing* step. For example the weighting of term for the previous sentence in section 3.2.1 is illustrated at Table 4.1.

Table 4.1 Weight term representation of document collection

TERM VECTOR MODEL BASED ON TF-IDF									
D_1: " Shipment of gold damaged in a fire" D_2: " Delivery of silver arrived in a silver truck" D3: " Shipment of gold arrived in a truck									
	Count, tfi						Weight, wi= tfi * IDFi		
Terms	D_1	D_2	D_3	df_i	D/df_i	IDF_i	D_1	D_2	D_3
a	1	1	1	3	3/3	0	0	0	0
arrived	0	1	1	2	3/2	0.1761	0	0.1761	0.1761
damaged	1	0	0	1	3/1	0.4771	0.4771	0	0
delivery	0	1	0	1	3/1	0.4771	0	0.4771	0
fire	1	0	0	1	3/1	0.4771	0.4771	0	0
gold	1	0	1	2	3/2	0.1761	0.1761	0	0.1761
in	1	1	1	3	3/3	0	0	0	0
of	1	1	1	3	3/3	0	0	0	0
silver	0	1	0	1	3/1	0.4771	0	0.4771	0
shipment	1	0	1	2	3/2	0.1761	0.1761	0	0.1761
truck	0	1	1	2	3/2	0.1761	0	0.1761	0.1761

The detail of Table 4.1represents the items below:

1. Columns 1: construct an index of terms from the documents, such as word: 'a', 'arrived', 'damaged', etc.
2. Column 2-4: determine the term count tf_i for each document D_j *(D_1, D_2, D_3)*
3. Columns 5-7: straightforwardly compute the document frequency df_i for each document and the total document D is equal to 3, so it can be obtained $IDF_i = \log(D/df_i)$.
4. Columns 8-10: compute the term weights each document. These columns can be described as a sparse matrix in which most entries are zero.

4.4 Dimension Reduction of Term-Document Matrix Representation

The volume of data or terms is very huge when documents are represented using the Vector Space Model. It also creates sparse matrix in which most entries are zero due to the elements of matrix contain only weight-term of presence term in document collection. Besides that, there is no information which words (terms) are important and have correlation each other. Many terms have the same proportion of weight-term

so that clustering based on similarity of weight-term cannot be related to the specific word (term). By decomposing the matrix with Singular Vector Decomposition (SVD) or Principal Component Analysis (PCA), the matrix dimension can be reduced by capturing the pattern of document collections. Hereby, the algorithm details of SVD and PCA:

A. SVD algorithm for decomposing document-term matrix A:

a. Compute its transpose A^T and A^TA
b. Get the eigenvalues of A^TA and arrange these in descending order, in the absolute sense. Then, compute the square roots of these to obtain the singular values of A
c. Construct a diagonal matrix $\sum$ by placing singular values in descending order along its diagonal. After that, compute its inverse, $\sum^{-1}$.
d. Use the ordered eigenvalues from step b and compute the eigenvectors of A^TA. Put these eigenvectors along the columns of V and count its transpose, V^T.
e. Compute membership matrix U as $U=AV\sum^{-1}$.

B. PCA algorithm for decomposing document-term matrix A:

Generally, this method seeks to get the eigenvector and eigenvalues from its covariance, as stated in details below:

a. Construct a $N \times d$ document-term matrix A, with one row vector A_n per data point
b. Then matrix A subtract mean is multiplied from each row vector A_n in A
c. Get the covariance matrix Y of A
d. Find eigenvector and eigenvalues of Y
e. The principal component is obtained from M eigenvectors with the largest eigenvalues

The PCA is known as applying Singular Vector Decomposition (SVD) on the covariance matrix. Here, the illustration of PCA for document-term matrix A is as shown below:

$A \rightarrow Y$, *where* $Y = A_i - \mu_j$
$Y \rightarrow Y^T$
$1/(n)\ Y^TY \rightarrow A$
$A \rightarrow U\sum V^T$

4.5 Mapping Term-Document Matrix to Document Matrix Using LSI Approach

SVD can be used to solve the low-rank matrix approximation of document-term matrices by applying the following three steps [56].

1. Construct SVD in form $A = U\sum V^T$ in a given matrix A ($m \times n$ dimension)
2. Get the derivation from $\sum$, the matrix $\sum_k$ formed by replacing with zeros the $m - k$ smallest singular values on the diagonal of $\sum$
3. Compute and generate to produce $\boldsymbol{A \approx A_k = U_k \sum_k V_k^T}$ as the rank-k of approximation to matrix A

Thus, the document-term matrix A ($m \times n$ dimension) representative is transformed to the document matrix V ($k \times n$ dimension) in LSI space. By using the properties of SVD matrix, this formulation $A = U\sum V^T$ can be derived as follows:

$$\boldsymbol{A = U \sum V^T}$$

$$\boldsymbol{\sum^{-1} \cdot U^T \cdot A = V^T}$$

Since $\cdot \boldsymbol{U^T = 1}$, then

$$\boldsymbol{V = A^T \cdot U \cdot \sum^{-1}}$$

In applying the low-rank matrix approximation can be stated in Equation (4.2):

$$\boldsymbol{V_k = A^T \cdot U_k \cdot \sum_k^{-1}} \quad (4.2)$$

The matrix V is used as document representative in concept space. It shows that there is matrix size reduction until k which represents the number of pattern for the terms inside the document. The size of k is less than or equal of the number of document which is based on the property of the decomposed low rank matrix. This matrix is also applied to PCA method in order to obtain document matrix in concept space.

4.6 Modified Fuzzy c-Means Algorithm

Basically, FCM clustering algorithm minimizes Fuzzy c-Means' objective function which measures the overall dissimilarity within clusters. The optimal clustering can be obtained by evaluating the following objective function as in Equation (3.4).

$$J'(U,V) = \sum_{i=1}^{c}\sum_{k=1}^{N}(\mu_{ki})^m D(x_k, v_i), \quad \sum_{i=1}^{c}\mu_{ik} = 1$$

where,

μ_{ik} = membership value of the word x_k in cluster

N = the total number of the word

c = number of the clusters

v_i = cluster center

D = distance

Thus, input of FCM

- Term weight data set $X = \{x_1, x_2, \ldots x_n\}$ as representative of A (after reduction)
- Number of clusters $1 < c < N$
- Weighting exponent $m > 1$
- Terminating tolerance $\boldsymbol{\varepsilon} > 0$
- Initialize the partition matrix $U^{(0)}$

Output of FCM

- Membership matrix $U\ (N \times c)$
- Prototype matrix $V\ (c \times n)$, n is the number terms/ words (pattern)

This is illustration of algorithms details which involves cosine similarity formulation.

Repeat for $l = 1, 2, \ldots$

Step 1: Compute the cluster prototype (means) v_i:

$$\overline{\boldsymbol{v}}_{\boldsymbol{i}} = \frac{\sum_{k=1}^{N}(\bar{u}_{ki})\frac{x_k}{\|x_k\|}}{\left\|\sum_{k=1}^{N}(\bar{u}_{ki})\frac{x_k}{\|x_k\|}\right\|}, \quad 1 \le i \le c$$

, Step 2: Compute the distances among clusters

$$D(x, v) = 1 - S_{cos}(x, v)$$

where,

D = dissimilarity between feature vector x and prototype v

S_{cos} = cosine similarity formula

Step 3: Update the partition matrix

$$U_{ki} = \left[\sum_{j=1}^{c}\left(\frac{D(x_k,\ \bar{v}_i)}{D(x_k,\bar{v}_j)}\right)^{\frac{1}{m-1}}\right]^{-1} = \left[\sum_{j=1}^{c}\left(\frac{1-S(x_k,v_i)}{1-S(x_k,v_j)}\right)^{\frac{1}{m-1}}\right]^{-1}$$

until

$$\left\|U^{(l)} - U^{(l-1)}\right\| < \varepsilon$$

where, U is partition matrix for evaluated data set.

Furthermore, the algorithm can be illustrated in a flowchart in Figure 4.3.

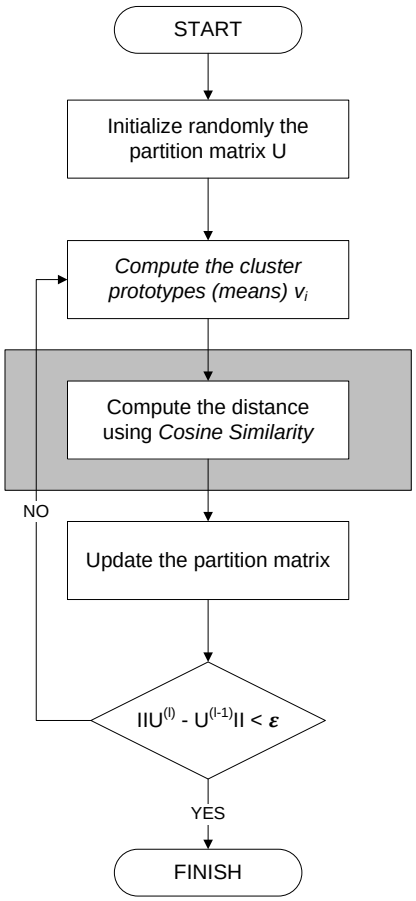


Figure 4.3 Flow chart of modified Fuzzy c-Means clustering

Figure 4.3 shows that there is a modification of Fuzzy c-Means algorithm in calculating the distance between centroid and data object using Cosine Similarity formula instead of Euclidean distance. This algorithm starts from initializing partition matrix randomly. Then, it computes the cluster prototype (means) as cluster center is obtained from the document in concept space within vector domain. After that, the partition matrix is updated repeatedly until small difference achieved. When updating cluster center, it needs to consider the distance formula.

4.7 Summary

The framework of this research has been described. The method and algorithm in details of each step are explained in this chapter. Representation of document in concept space can be derived from matrix decomposition. Then the cosine similarity as the replacement of Euclidean distance can be embedded to objective function of Fuzzy c-Means algorithm.

CHAPTER 5

EXPERIMENTAL RESULT AND DISCUSSION

5.1 Introduction

In this chapter, we present our result and analysis of the experiments discussed in chapter four. The experiments have been tested in Matlab7 that runs under Windows XP SP2 operating system. The first part of this chapter describes the data sets used for this research and describes the result of the *preprocessing* step. We use a Term-Document Matrix Generation (TMG) toolbox to generate and represent the document text into a term-document matrix. The clustering algorithm is then applied to document-term matrix in concept space. Finally, the performance cluster quality is analyzed for document categorization.

5.2 Data Preparation

The experiments are conducted using three data sets with various topics and data volume. The data sets are 20 News Group, Classic3, and YahooK1. By observing the data sets' categories, it uses various data sets' complexity natures. There are two kinds of data volume in a data set, i.e. balanced data volume and unbalanced data volume. The first condition is defined for each class (topic) which has relatively similar number of documents, i.e. data set 'News Group' (binary2, multi5, and multi10) and data set 'Classic3'. Then, the second condition is defined for each class that has extremely unbalanced number of document, i.e. data set YahooK1. Besides, the structure of class (category or topic) also needs to be considered. The class structure can be divided into three attributes which are well separated, overlapping, and highly overlapping. The well separated class is defined to all classes with different topics, i.e. multi5, classic3, and YahooK1. In contrast, the overlapping class is defined to classes that have several similar topics, i.e. multi10. And, the highly overlapping class is defined to all classes that consist of similar (overlapping) topics,

i.e. binary2. Overlapping categories or classes indicate overlapping features or words [57]. A brief description of the three data sets is mentioned in Table 5.1.

The first data set is taken from http://kdd.ics.uci.edu/databases/20newsgroups.html and contains 20000 messages taken from 20 News Groups. The messages (articles) are typical postings and have headers which include subject lines, signature files, and quoted portions of other articles. These parts are presented in APPENDIX B. The data consist of short news with various topics and data volume of subsets. As shown in Table 5.1, there are three subsets (binary2, multi5, and multi10). They have various topics per subset and class size. Each subset consists of 500 documents. The data set 'binary2' consists of two topics which are balanced volumes with 250 documents of 'talk.politcs.mideast' and 250 documents of 'talk.politics.misc'. The two topics are both about politics, so there is overlapping words in the content of document. This subset has a total of 16940 terms. However, another dataset, Multi5 consists of five well separated topics in each 100 documents. The last data set, Multi10 consists of ten topics but there are several overlapping topics.

Another data set used for this experiment is Classic3 taken from ftp://ftp.cs.cornell.edu/pub/smart. The data set contains 1446 information retrieval abstracts from the Cisi collection, 1398 aerospace system abstracts from the Cranfield collection, and 1031 medical abstracts from Medline collection. There are overall 3875 documents and 160034 terms.

The last data set is YahooK1 which is taken from ftp://ftp.cs.umn.edu/dept/users/boley/pddpdata/doc-K and contains 2340 reuters news articles from Yahoo in 1997. There are 6 categories which consist of 494 from Health, 1389 from Entertainment, 141 from Sports, 114 from Politics, 60 from Technology and 142 from Business. The documents are in html format file and have implicit topics inside them. Furthermore, as illustration, the document layout for the data sets: 20News Group, Classic3 and YahooK1 shown in APPENDIX B, and briefly the data details are shown in Table 5.1. The original document of data set'Classic3' is shown in Figure 5.1.

Table 5.1 Data set details

Dataset	Topics	Total Docs	Total Terms	Explanation
Binary2	talk.politics.mideast talk.politics.misc	500	16940	Balanced and highly overlapping
Multi5	comp.graphics rec.motorcycles rec.sport.baseball sci.space talk.politics.mideast	500	18056	Balanced and well separated
Multi10	alt.atheism comp.sys.mac.hardware misc.forsale rec.autos rec.sport.hockey sci.crypt sci.electronics sci.med sci.space talk.politics.guns	500	18565	Balanced and overlapping
Classic3	information retrieval (cisi) aerospace (cranfield) medical (medline)	3875	160034	Balanced and well separated
YahooK1	Health, Entertainment, Sports, Politics, Technology, Bussiness	2340	38807	Unbalanced and well separated

```
.I 1
18 Editions of the Dewey Decimal Classifications
.A
Comaromi, J.P.
.W
  The present study is a history of the DEWEY Decimal Classification.  The first edition of the DDC
was published in 1876, the eighteenth edition in 1971, and future editions will continue to appear
as needed.  In spite of the DDC's long and healthy life, however, its full story has never
been told.  There have been biographies of Dewey that briefly describe his system, but this is the
first
attempt to provide a detailed history of the work that more than any other has spurred the growth
of
librarianship in this country and abroad.
.X
1       2       1
1       2       1
1       2       1
1       2       1
1       2       1
556     2       1
.I 2
</TEXT>
```

Figure 5.1 An original document from data set ‘Classic3’

Initially, *preprocessing* step is applied which parses the contents of the document with the specific character ‘#’ and inserted into recognized delimiter among terms or

words. Beside *parsing terms* in document collection, *stemming* and *removing stop word* are also applied. The document result after *preprocessing* is shown in Figure 5.2.

edit# dewei# decim# classif#camaromi#present# studi# histori# dewei# decim# classif# edit# ddc#publish# eighteenth# edit# future# edit# continu# spite#ddc# healthi# full# stori#told# biographi# dewei# briefi# describ# system#attempt# provid #detail# histori# spur# growth#librarianship# country# abroad#

Figure 5.2 A document 'Classic3' after *preprocessing*

The thresholds are also given and the details shown in Table 5.2. As an illustration, data set 'binary2' is given a minimum threshold length of 3 (three) characters and a maximum threshold length of 30 characters in each word (term). The minimum of word frequency per document is one word known as local frequency. Another threshold is global frequency which shows the number of words in the whole document collection. To define data of thresholds, we refer to information theory which mentioned that the high frequency of words in a document indicates that the words are not important.

Table 5.2 Data of thresholds

Data set	Total docs	Length		Local frequency		Global frequency	
		Min	Max	min	Max	Min	Max
Binary2	500	3	30	1	250	3	500
Multi5	500	3	30	1	250	3	500
Multi10	500	3	30	1	250	3	500
Classic3	3875	3	30	1	1500	3	3000
YahooK1	2340	3	30	1	1000	3	2000

Applying threshold in *preprocessing* stage reduces number of terms. The details of reduction are shown in Table 5.3.

Table 5.3 Data reduction details after *preprocessing*

Data Sets	Total Docs	Total Terms (before)	Total reduction					Total Terms (after)	
			Stop Words	Stem-ming	Term Length	Threshold		in number	in %
						Global	Local		
Binary2	500	16940	428	4306	208	7193	1	4804	71.64
Multi5	500	18056	427	3961	347	8939	4	4378	75.75
Multi10	500	18565	433	4176	420	9205	76	4255	77.08
Classic	3875	160034	431	7015	266	9535	137441	5346	96.66
YahooK1	2340	38807	440	11431	275	14237	0	12424	67.99

In Table 5.3 terms of data sets have been reduced. As an illustration, data set 'binary2' has been reduced, the details are as follows: 428 terms of stop word removal, 4306 terms of stemming, 208 terms of term-length which less than 30 characters, 7193 terms and 1 term of global and local frequent threshold. Therefore, the current number of terms for 'binary2' after *preprocessing* is 4804 terms which is derived from subtraction of total reduction to total terms. It means that there is data set reduction to 71.64% from the original data.

The next step is representing it into Vector Space Model of document using TF-IDF weight-term formulation as shown in Table 5.4. Term 'abstract' has proportion as 0.66, 0.34, 0.20 and 0.27 in the document collections: D006, D037, D043 and D045. Also, term 'abstractor' has weighted term as 0.23 in document number six (D006).

Table 5.4 Vector Space Model

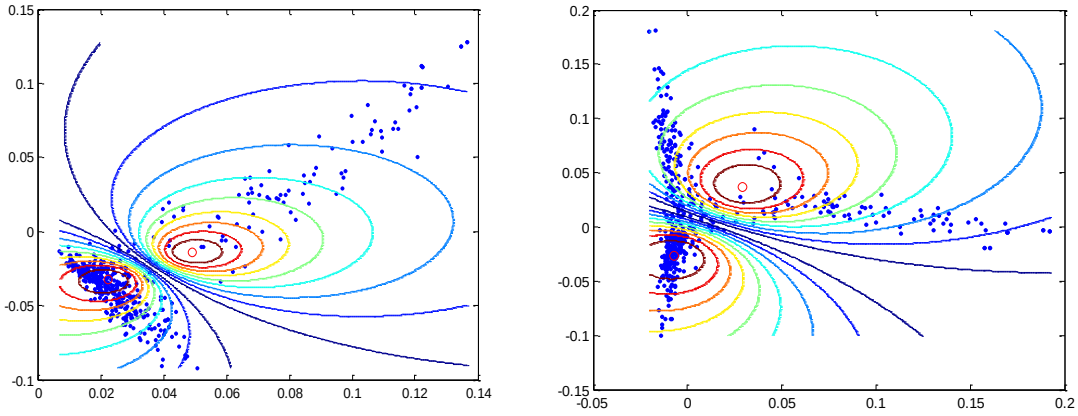
	D001	..	D004	..	D006	..	D009	D010	..	D019	..	D037	..	D043	..	D045
abnorm	0		0		0		0	0		0		0		0		0
abolish	0		0		0		0	0		0		0		0		0
abort	0		0		0		0	0		0		0		0		0
absenc	0		0		0		0	0		0		0		0		0
absent	0		0		0		0	0		0		0		0		0
absolute	0		0		0		0	0		0		0		0		0
absorb	0		0		0		0	0		0		0		0		0
absorpt	0		0		0		0	0		0		0		0		0
abstract	0		0		0.66		0	0		0		0.34		0.20		0.27
abstractor	0		0		0.23		0	0		0		0		0		0
academ	0		0.41		0		0	0		0		0		0		0
acceler	0		0		0		0	0		0		0		0		0
accept	0		0		0		0	0		0		0		0		0
access	0		0		0		0.27	0.42		0		0		0		0
accessory	0		0		0		0	0		0		0		0		0
accommod	0		0		0		0	0		0		0		0		0
accompani	0		0		0		0	0		0		0		0		0
accomplish	0		0		0		0	0		0		0		0		0
accord	0		0		0		0	0		0		0		0		0
account	0		0		0		0	0		0.18		0		0		0

In document collection representation, the sparse matrices created have dimensions which are still considered as large. However by applying the LSI, further reductions can still be made.

5.3 Result of Fuzzy c-Means Clustering Algorithm

The decomposed matrix is also intended to map the document into concept space. After the document is mapped into concept space, it is then subjected to the clustering algorithm itself using 'Cosine similarity' in the objective function. The clustering algorithm involves some parameters such as: fuzziness (m=1.1), error rate ($\boldsymbol{\varepsilon}$=0.0001) and the cluster number which is same as the number of topics or categories.

In Fuzzy c-Means algorithm, we initialize the memberships randomly and it will terminate until the maximum change in membership less than error rate ($\boldsymbol{\varepsilon}$). The illustration of document clustering for data set 'Binary2' using SVD and PCA are plotted in 2-dimension as shown in Figure 5.3.



(a) SVD (b) PCA

Figure 5.3 Illustration of Fuzzy c-Means clustering algorithm by the data set 'Binary2'

In Figure 5.3, the dot ('.') remarks the data points, the 'o' is cluster center which is the weighted mean of data. The Fuzzy c-Means algorithm can only detect clusters with circle shape as the characteristic of this algorithm. The contour line depicts the membership of clustering result and bounds data (term-documents) in a cluster as illustration of members inside it. There is difference of data distribution between SVD and PCA. This difference influences the grouping of membership in each cluster. Furthermore, it indicates the effect to their performance of cluster quality. For all data sets is shown in Appendix D.

The output of fuzzy clustering algorithm is a membership degree of document for each cluster. Therefore, we use the Fuzzy Prediction Method (PM 1) to select the maximum value of membership degree to define which document belongs to a certain cluster number. Therefore, we summarize the number of document clustering results for all data sets into specific cluster as shown in Table 5.5.

Table 5.5 The number of document clustering result

Data set	Cluster number	Number of documents	
		SVD- Cosine	PCA-Cosine
Binary2	1	250	250
	2	250	250
Multi5	1	95	100
	2	99	99
	3	98	95
	4	101	99
	5	107	107
Multi10	1	55	54
	2	58	59
	3	42	54
	4	42	45
	5	63	54
	6	47	44
	7	57	51
	8	55	57
	9	39	39
	10	42	43
Classic3	1	1396	1458
	2	1117	1414
	3	1362	1003
YahooK1	1	583	271
	2	473	306
	3	261	478
	4	331	532
	5	278	475
	6	414	278

Table 5.5 shows the number of documents which has been grouped into the respective clusters for all data sets. The cluster number refers to the number of category (class). There are different number of documents between using SVD and PCA for data set Multi5, Multi10, Classic3, and YahooK1. Besides the number of document in each cluster, some mistakes in clustering are analyzed using confusion matrix which will be discussed on the next section. Furthermore, the details of document clustering results are shown in Appendix E.

5.3 Evaluation Measurement for Document Clustering

It is important to know the performance quality of document clustering. In the clustering approach, the performance normally depends on internal and external quality based on the similarity. Three measures that are widely used are entropy, F-measure, and accuracy. These measures use labeled test document collection. There is a comparison between the cluster results and labeled classes. They measure the degree to which documents from the same classes are assigned to the same cluster.

5.3.1 Entropy

In order to investigate the internal quality of clustering, the evaluation measurement uses entropy. The entropy measures homogeneity in clustering and the formula is stated below:

$$E_j = -\sum_i P(i,j) \cdot \log P(i,j) \tag{5.1}$$

where, *P(i, j)* is the probability that a document has class label *i* and is assigned to cluster *j*. Thus, the total entropy of clusters is obtained by summing the entropies of each cluster weighted by the size of each cluster as denoted by the following formula:

$$E = \sum_j \frac{n_j}{n} E_j \tag{5.2}$$

where, n_j is size of cluster *j* and n is total document number in the corpus. A lower entropy value signifies a higher quality of cluster and alternatively a higher homogeneity [58].

5.3.2 F-Measure

The F-measure includes precision and recall [58]. This basic idea is from the information retrieval concept. In this measure, each cluster is as if it is the result of query and each class as if it is the desired set of documents for the query. Precision and recall for each cluster *j* and class *i* are denoted as follows:

$$Recall(i,j) = \frac{n_{ij}}{n_i}$$

$$Precision(i,j) = \frac{n_{ij}}{n_j} \quad (5.3)$$

where, n_{ij} is the number of documents with class label i in cluster j, n_i is the number of documents with class label i and n_j is the number of documents in cluster j. Consequently the F-measure for cluster j and class i can be obtained as below:

$$F(i,j) = \frac{(2 * Recall(i,j) * Presicion(i,j))}{Recall(i,j) + Precision(i,j)} \quad (5.4)$$

In the hierarchical clustering, the F-measure of any class is the maximum value it achieves at any node in the hierarchy tree. An overall value for the F-measure is the weighted average f all values for the F-measure:

$$F = \sum_i \frac{n_i}{n} \max_j F(i,j) \quad (5.5)$$

The value of the F-measure ranges from [0..1] and a larger value corresponds to a better clustering quality.

5.3.3 Accuracy

The accuracy of a cluster $\hat{C}_j$ in class C_i is formulated as in Equation (5.6) [58]:

$$A(\hat{C}_j) = \max_{1=1}^{k} \frac{|C_i \cap \hat{C}_j|}{|\hat{C}_j|} \quad (5.6)$$

The accuracy of k-clustering $C_1 \ldots C_k$ is the weighted sum of accuracies. The accuracy in hierarchical clustering is the maximum accuracy of any choice of k nodes that group C. The range of accuracy value is between 0 and 1 or in percentage (0-100%). The higher accuracy value, the better performance of clustering is. Accuracy, which has been used as a measure of performance in supervised learning, has been used in clustering by using confusion matrix.

5.4 Performance Evaluation of Document Clustering

All the experiments are carried out to analyze the product clusters from Fuzzy c-Means clustering with LSI approach in reducing of matrix dimension size. We have used membership degree as a basis for determining document into specific cluster number.

There is a significant reduction of matrix dimension which is represent the document in concept space with size of *k*-rank by total document. The terms are reduced to small values which refer to *k*-rank. That means there are *k* patterns in each document collection. To find out the effect of *k*-rank to quality of cluster, we apply various *k*-ranks to cluster of mapped documents. The number of *k*-rank effects to decompose term-document matrix. The different *k*-rank makes different factorial of matrix elements which representative of weighting term in document collections. The rank is used *k*=2,3,4,..,*n* where *n* is total document and is illustrated in Figure 5.4. This figure depicts the performance of applying Singular Vector Decomposition (SVD) for clustering data set 'multi5'. It shows that the *k*-rank=13 is at the optimum condition with good performance (high f-measure, and low entropy).

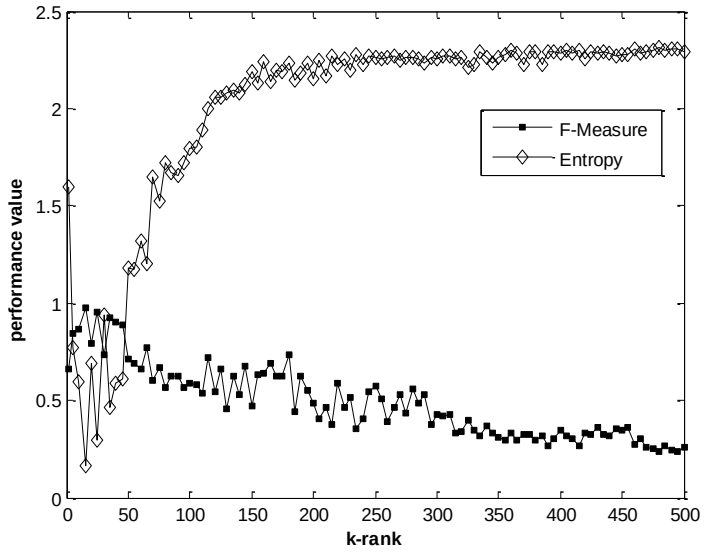


Figure 5.4 Performance evaluation of clustering 'Multi5' with various *k*-ranks using SVD

The PCA method is also implemented for clustering with certainty *k*-rank. Similar to SVD method, it is used various *k*-rank =2,3,4,…*n* (total document) and the result is

depicted as in Figure 5.5. This figure shows that clustering using low rank (*k*-rank=7) can obtain the best performance.

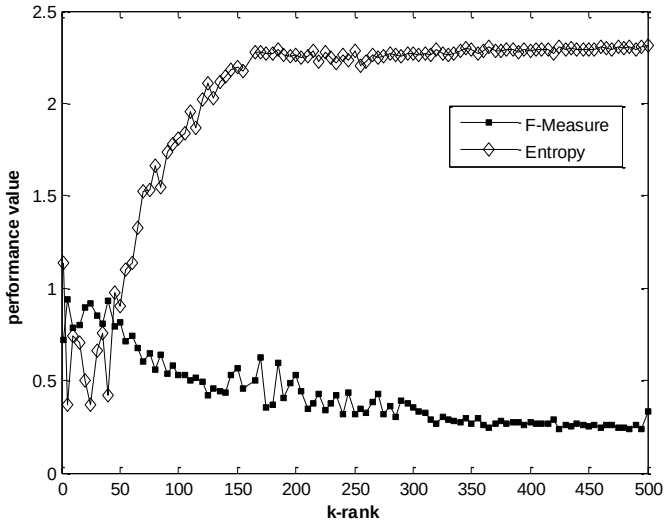


Figure 5.5 Performance evaluation of clustering 'Multi5'with various *k*-ranks using PCA

Thus, we apply clustering to all data sets and get the optimum *k*-rank for each data set as it is shown in Table 5.6. It shows that there is a correlation between total cluster and data volume of *k*-rank. The optimum *k*-rank tends to low its number. In summary, the optimum *k*-rank is in interval [2..50] for all data set in this research. Thus, Appendix C illustrates the correlation between *k*-rank and the performance of clustering result for all data sets.

Table 5.6 Optimum *k*-rank of data sets

Data sets	Total Document	Total Cluster	*k*-rank (SVD)	*k*-rank (PCA)
Binary2	500	2	12	4
Multi5	500	5	13	7
Multi10	500	10	44	38
Classic	3875	3	12	7
YahooK1	2340	6	20	15

The experiments use data sets with various class sizes. To know the correlation between class size and number of optimum *k*-rank, we plot as in Figure 5.6. A large class size tends to have the small number of *k*-rank.

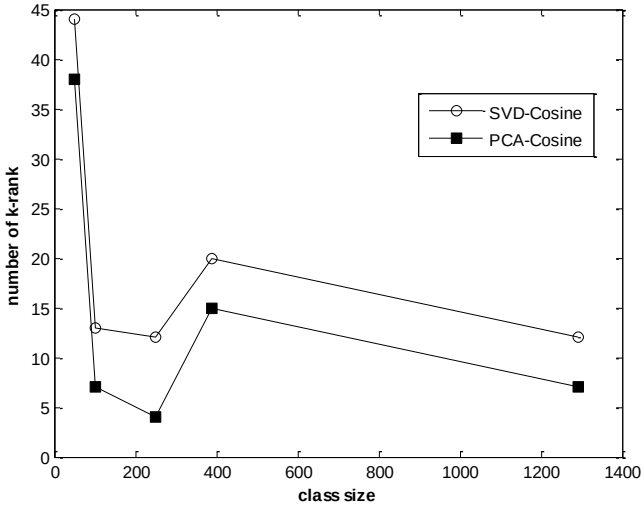


Figure 5.6 Class size versus number of *k*-rank

Another hand, the correlation between the number of cluster and the number of optimum *k*-rank for all data sets of this research can be plotted in Figure 5.7. This figure shows that the cluster number increases, the optimum k-rank also increases.

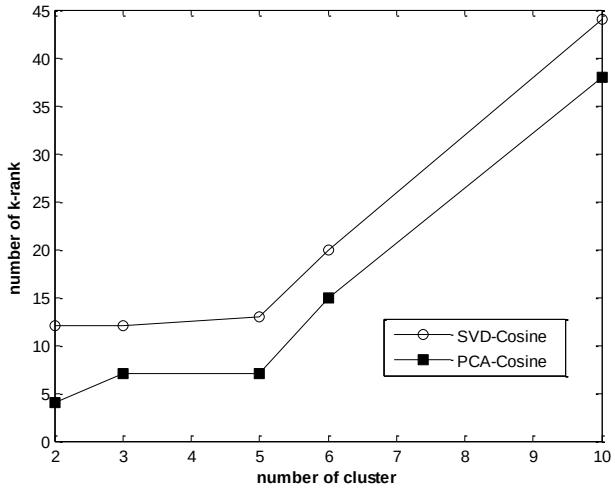


Figure 5.7 Cluster number versus number of *k*-rank

Furthermore, by using the optimum *k*-rank we reach the best performance of document clustering for each data set. The standard measurements of evaluation are precision, recall, f-measure, and entropy. However, the information of these measurements for the other three algorithms (FCCM, FSKWIC, and Fuzzy CoDoK) is unavailable. So we compare the proposed method to unapplied LSI method.

Precision is a tool to identify the ratio of relevant retrieved word or term against retrieved word or term in the document collection. For this experiment we compared the precision of all datasets to a control data set which is not subjected to the LSI. In addition, two distance measures, namely the Euclidean and the Cosine measurement are used. The numerical result in Table 5.7 shows that the matrix decomposition (SVD, PCA) combined with Cosine similarity gives the highest precision compared to other methods. The high precision indicates the high performance quality of clustering [58].

Table 5.7 Precision of document clustering

Data sets	No LSI	SVD Euc	SVD Cosine	PCA Euc	PCA Cosine
Binary2	0.58	0.898	0.920	0.862	0.924
Multi5	0.742	0.886	0.978	0.918	0.972
Multi10	0.574	0.736	0.830	0.718	0.856
Classic3	0.735	0.798	0.955	0.767	0.962
YahooK1	0.668	0.86	0.876	0.827	0.869

Similarly, another experiment was conducted to obtain the recall rate. This rate is ratio between relevant retrieved word or term against the relevant word or term. Comparison of various methods is applied for this measure as shown in Table 5.8. The proposed method shows the highest recall. However, this condition is not occurred to data set 'YahooK1' which LSI with Euclidean (SVD-Euclidean and PCA-Euclidean) method is superior. It also gained the worst results in the application of the proposed method compared to the other data sets. Highly unbalanced volume of this data set affects the recall rate but it is still shows a satisfactory performance [58].

Table 5.8 Recall of document clustering

Data sets	No LSI	SVD Euclidean	SVD Cosine	PCA Euclidean	PCA Cosine
Binary2	0.580	0.900	0.920	0.882	0.927
Multi5	0.658	0.903	0.979	0.930	0.973
Multi10	0.523	0.779	0.830	0.780	0.859
Classic3	0.620	0.825	0.949	0.819	0.965
YahooK1	0.722	0.874	0.797	0.840	0.811

F-measure as external performance evaluation which combines from precision and recall measurements has the numerical result as it is illustrated in Figure 5.8.

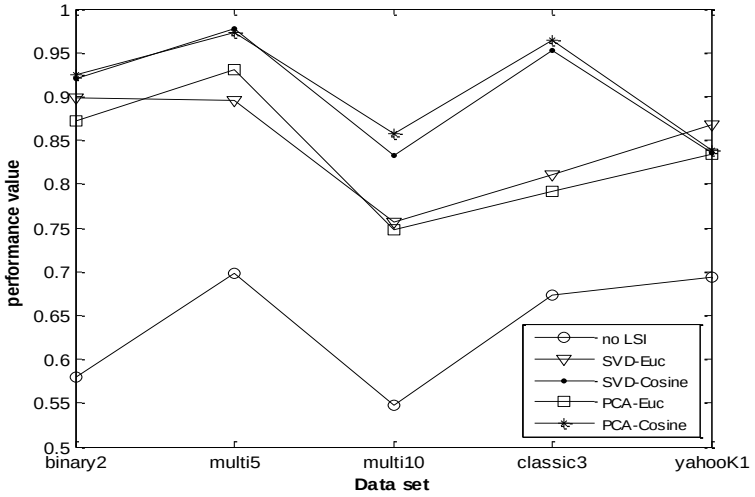


Figure 5.8 Comparison F-measure for all data sets

In Figure 5.8, the performance evaluation using f-measure shows that the proposed method (SVD-Cosine and PCA-Cosine) consistently outperforms the other algorithms. The average f-measure of the proposed method for all data sets are the highest as it is shown in Table 5.9. However, when we applied to data set 'YahooK1', the F-measure of SVD-Euclidean and PCA-Euclidean is superior to the proposed method. This matter is caused that the ratio of relevant retrieved words against the relevant words (recall) of the proposed method is lower than those methods.

Table 5.9 F-measure of document clustering for all data sets

Data sets	No LSI	SVD Euclidean	SVD Cosine	PCA Euclidean	PCA Cosine
Binary2	0.58	0.899	0.92	0.872	0.925
Multi5	0.698	0.895	0.978	0.93	0.973
Multi10	0.547	0.757	0.832	0.748	0.858
Classic3	0.673	0.811	0.952	0.792	0.964
YahooK1	0.694	0.867	0.835	0.834	0.839

Another performance indicator is entropy which is used to measure the internal quality of a given cluster. Figure 5.9 compares the quality of all data sets with various methods. The result shows that the performance of SVD-Cosine and PCA-Cosine is always superior to the method without LSI. The entropy value which is close to 0

indicates good performance of clustering [58]. The lowest entropy occurred on data set ‘multi5’, which has well-separated topics. In contrast, the highest entropy is data set ‘multi10’. This data set has overlapping topic in the document collection and has small class size. Furthermore, the entropy details of all data sets are shown in Table 5.10.

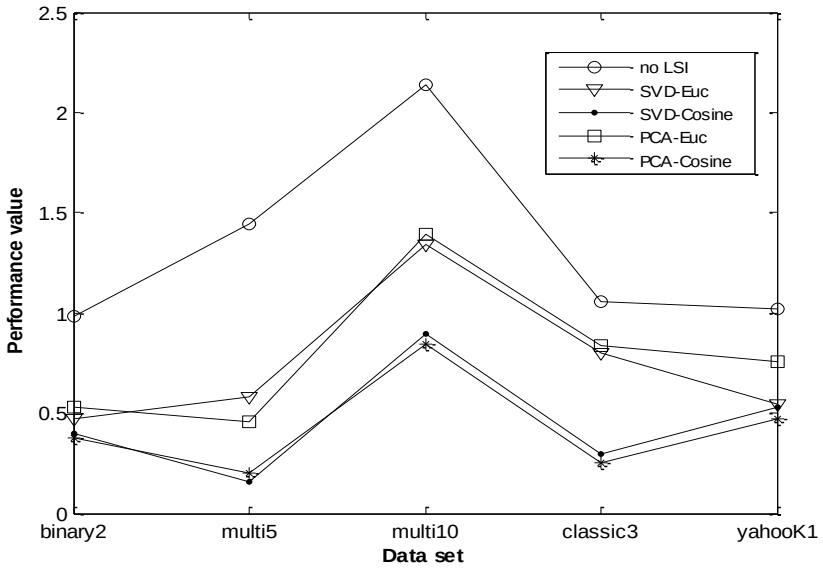


Figure 5.9 Entropy of document clustering

Table 5.10 The data details of entropy of document clustering for all data sets

Data sets	No LSI	SVD Euclidean	SVD Cosine	PCA Euclidean	PCA Cosine
Binary2	0.982	0.468	0.402	0.527	0.375
Multi5	1.445	0.578	0.157	0.459	0.200
Multi10	2.139	1.341	0.898	1.391	0.843
Classic3	1.054	0.799	0.299	0.835	0.249
YahooK1	1.0213	0.543	0.528	0.760	0.468

5.4.1 Confusion Matrix

A confusion matrix is a tool in which a predicted class represents the instances in an actual class. It is easy to see if the system confuses the two classes, such as mislabeling one to another. Partitioning the entire document collection into several clusters by modified Fuzzy c-Means generates the confusion matrix given. The entry *ij* is the number of documents that belongs to cluster *i* and document collection *j*.

Since there is no connection between cluster *i* and document collection *j*, then the cluster cannot be defined by the confusion matrix in which it should be in diagonal matrix unless the rows or columns of matrix (suitably permuted). A good clustering produces a single dominant (the most) entry in each row in confusion matrix. Therefore, the rest of the entries in a row are defined as misclassifications. The number of misclassified documents is the sum of all misclassification in a confusion matrix.[59] Table 5.11 until Table 5.15 shows the confusion matrix using the Singular Vector Decomposition approach. The misclassified document details are shown in the non highlighted numbers in the cell.

Table 5.11 Confusion Matrix for data set 'Binary2' using SVD

	DC01	DC02
clust# 1	230	20
clust# 2	20	230

Clustering for the data set 'binary2' in Table 5.11 shows that there are 40 misclassified documents with details: 20 documents in cluster-1 and 20 documents in cluster-2. DC01 and DC02 are considered to document collection with topic 'politic.mideast' and 'politic.miscelaneous'. The confusion matrix shows that there is 8% misclassification of the entire document collections for categorization.

Another data set from the 20News Group is data set 'multi5' also occurred 13 misclassified documents (2.6% of the entire document collections) as shown in Table 5.12. The document collections in a certain topic are denoted in DC, e.g. DC01(comp.graphics), DC02(motorcycles), DC03(sport.baseball), DC04(sci.space), and DC05(politics.mideast). The details misclassified are two documents of cluster-2 , one document of cluster-3, one document of cluster-4, and seven documents of cluster-5.

Table 5.12 Confusion Matrix for data set 'Multi5' using SVD

	DC01	DC02	DC03	DC04	DC05
clust #1	0	0	95	0	0
clust #2	0	97	2	0	0
clust #3	0	0	1	97	0
clust #4	0	1	0	0	100
clust #5	100	2	2	3	0

The last data set of the same resource is data set ‘multi10’ in Table 5.13 which has 85 misclassified documents (17% of the entire document collections).

Table 5.13 Confusion Matrix for data set ‘Multi10’ using SVD

	DC01	DC02	DC03	DC04	DC05	DC06	DC07	DC08	DC09	DC10
clust#1	0	0	0	0	3	48	3	1	0	0
clust#2	1	3	45	0	4	0	1	2	2	0
clust#3	0	0	1	0	40	0	0	1	0	0
clust#4	2	0	0	0	1	0	24	6	9	0
clust#5	2	1	0	49	2	0	7	2	0	0
clust#6	1	0	2	0	0	0	7	36	1	0
clust#7	1	1	2	0	0	0	1	0	2	50
clust#8	0	45	0	1	0	1	6	2	0	0
clust#9	1	0	0	0	0	1	1	0	36	0
clust#10	42	0	0	0	0	0	0	0	0	0

In clustering document for data set ‘Classic3’, the good clustering is obtained in highlighted numbers of the table as shown in Table 5.14. There are various misclassified documents in each cluster which is show in non-highlighted numbers with total are 183 documents (4.72%).

Table 5.14 Confusion Matrix for data set ‘Classic3’ using SVD

	DC01	DC02	DC03
Cluster #1	30	1347	19
Cluster #2	74	40	1003
Cluster #3	1342	11	9

The last applying SVD method is to cluster document of data set ‘YahooK1’ and it is found 372 misclassified documents (15.89%). The details are 2 documents of cluster-1, 9 documents of cluster-2, 23 documents of cluster-3, 190 documents of cluster-4, 148 documents of cluster-5 and 372 documents of cluster-5 as shown in Table 5.15

Table 5.15 Confusion Matrix for data set 'YahooK1' using SVD

	DC01	DC02	DC03	DC04	DC05	DC06
clust#1	0	581	0	0	0	2
clust#2	64	7	0	0	0	2
clust#3	22	238	0	0	0	1
clust#4	4	67	141	112	0	7
clust#5	4	82	0	2	60	30
clust#6	0	414	0	0	0	0

Similarly another method, PCA is applied to cluster document of all data sets. The mistakes are also generated using Confusion matrix and as a result, the misclassified documents are shown in Table 5.16.

Table 5.16 The details of misclassified document clustering using PCA

Data sets	Number of misclassified documents in each cluster	Percentage of misclassified documents
Binary2	clust#1(29), clust#2(9)	7.6%
Multi5	clust#1(2), clust#3(1), clust#4(2), clust#5(9)	2.8%
Multi10	clust#1(7), clust#2(13), clust#3(5), clust#4(3), clust#5(7), clust#6(1), clust#7(14), clust#8(11), clust#9(7), clust#10(4)	14.4%
Classic3	clust#1(75), clust#2(29), clust#3(34)	3.56%
YahooK1	Clust#1(131), clust#2(171), clust#4(8), clust#5(2), clust#6(20)	14.19%

5.4.2 Comparison of accuracy to FCCM, FSKWIC, and Fuzzy CoDoK

By using confusion matrix, the number of misclassified documents can be known. Accuracy rate is the ratio between the number of correct classified documents and the number of all documents.

This experiment compared the proposed method with other algorithms: Fuzzy Clustering for Categorical Multivariate Data (FCCM), Fuzzy Simultaneous Keyword Identification, Clustering (FSKWIC) and Fuzzy Co-clustering of Documents and Keywords (Fuzzy CoDoK) [25, 60, 61]. The results are shown in Table 5.17, where it can be seen that both of our algorithms (FCM-SVD and FCM-PCA) outperform the other algorithms. High accuracy of more than 80% can be achieved by both algorithms.

Table 5.17 Comparison of accuracy using different algorithms

Data set	FCCM	FSKWIC	Fuzzy CoDoK	FCM noLSI	FCM SVD	FCM PCA
Binary2	71.99%	70.57%	70.55%	58%	92%	92.4%
Multi5	22.75%	62.71%	66.35%	54.6%	97.8%	97.2%
Multi10	22.42%	41.59%	49.52%	43.2%	83%	85.6%
Classic3	89.95%	88.97%	96.74%	62.78%	95.3%	96.46%
YahooK1	76.99%	65.95%	83.94%	74.78%	84.1%	85.81%

5.5 Summary

In this chapter, the results of methodology implementation are shown. There are significant reductions of terms either before representing document in matrix space and during term-document matrix decomposition. Before representing a document, the reduction of terms in *preprocessing* step is over 50% of original data. Thus, during decomposing matrix, the reduction of dimension size for term-document matrix is as the *k*-rank by the number of document. When constructing matrix decomposition, the *k*-rank gives impact to the performance quality of clustering (precision, recall, f-measure and entropy) is good, i.e. always greater than 70%. Moreover, the accuracy of the proposed method is also the highest among three other methods (FCCM, Fuzzy CoDoK, and FSKWIC).

CHAPTER 6

CONCLUSION AND FUTURE WORK

The research conducted aims to cluster documents whereby this research has developed a document clustering method using Latent Semantic Index (LSI) approach to improve its performance particularly in the small class size. The LSI approach is a method that is able to find the pattern of terms inside a document collection. Comparison has been undertaken based on the correct classification in a certain category. This research had uses various topics and data volumes to know the overall performance of document clustering. This research had also compared the method accuracy to other algorithms in the fuzzy domain.

The following chapter summarizes the central finding of this thesis and its contributions. Finally, this thesis can be used as challenge for associated future research.

6.1 Conclusion

Based on the various literatures reviewed on news category and the experiments results of this study, the usage of the proposed method to cluster document has shown an overall good performance.

There are five different structures of data sets used on this research which are ‘binary2’, ‘multi5’, ‘multi10’, ‘Classic3’, and YahooK1. The biggest class size for the data set is the ‘classic3’ while the smallest size is data set ‘multi10’. These data sets can be denoted in descending order of size as Classic3 > YahooK1 > binary2 > multi 5 > multi10. The data is then tested by considering not only to its class size, but also to the content structure, either by overlapping or separated topics.

A document is composed of many terms and words. In reality, a document is often incomplete, noisy, and inconsistent. Therefore, the first step of document clustering is

preprocessing, which includes *case folding*, *parsing*, *removing stop word*, and *stemming*. This step is done in order to remove meaningless information such as: numbering, article, preposition, common word, etc. from the original document. The *threshold* of frequent terms in *preprocessing* is intended to keep the correlation and consistency among the terms used inside document collections. As result from the *preprocessing* step, the data volume from the original document has significantly been reduced.

In document representation, the weight-term is counted using TF-IDF formula into vector space model. The numbers existed represents the available weight-term. Therefore, they perform the sparse matrix which has the zero entries bigger than non zero. Apart from that, there is a possibility of some words ended up having the same meaning (synonymy) and one word has various meanings (polysemy) in the document collections. Some words exist in documents concurrently interpreted as the same meaning. As for that reason, the LSI approach is applied to avoid any synonymy and polysemy theoretically. The LSI approach is a method to get the pattern in document collection which has been applied for this research, either SVD or PCA. The number of pattern is notated as *k*-rank. Having various *k*-ranks, we selected one of them at the best performance and as a result, our experiment obtained low value ($k < 50$). From this matrix decomposition, the document can be mapped into concept space based on the matrix decomposition's properties. The mapped document is then applied to a fuzzy clustering algorithm.

Fuzzy c-Means is one of fuzzy clustering algorithms recognizing spherical clouds of points in a *p*-dimensional space with the same size. This algorithm is basically aims to minimize the objective function related to the dissimilarity or distance formula. By using variant of the distance formula, Cosine Similarity replaces the original formula, Euclidean distance, and it is embedded to the algorithm.

The proposed method has been implemented and the result shows that the overall performance achieves more than 0.9 of external quality and below than 0.5 of internal quality for all kinds of data sets. The method has the highest average accuracy compared to the other three algorithms which uses key-word based; Fuzzy Categorical Multivariate Data (FCCM), Fuzzy Simultaneous Keyword Identification, Clustering (FSKWIC) and Fuzzy Co-clustering of Documents and Keywords (Fuzzy

CoDoK). The average accuracy it manages to achieve is higher than 90%. However, the other three algorithms results in lower accuracy when they are applied to small class size of data sets, such as data set ‘multi10’.

6.2 Future Work

This research uses document clustering for categorization. The clustered document is under one category. All of the results are returned to a value that refers to the prediction method (PM1). The reason is based on the data set form. We can enhance it using different form of data set that covers multi-category in one document.

In decomposition matrix, the *k*-rank is a representative of the number of pattern inside a data collection. It is also applied in order to map document in concept space. There is a reduction of document-term matrix dimension as *k*-rank by number of document. By substituting *k*-rank with several numbers, the document-term matrix is decomposed using SVD or PCA. It is repeatedly applied to the clustering algorithm in order to achieve the best performance. Therefore, it consumes large computation time. Because of this problem, another method can be applied to define the *k*-rank before it is embedded to clustering algorithm. With selected *k*-rank, the iteration is to find the cluster center and define the membership degree which is then can be reduced.

REFERENCES

[1] P.-N. Tan, M. Steinbach, and V. Kumar, *Introduction to Data Mining*, Pearson International ed.: Pearson Education, Inc., 2006.

[2] M. A. Hearst and J.O.Pedersen, "Reexamining the cluster hypothesis," in *Proceeding of SIGIR '96*, Zurich, Switzerland, 1996, pp. 76-84.

[3] N. Jardine and C. J. v. Rijsbergen, *The Use of Hierarchical Clustering in Information Retrieval* vol. 7, 1971.

[4] J. He, A.-H. Tan, C.-L. Tan, and S.-Y. Sung, "On Quantitative Evaluation of Clustering Systems," in *Clustering and Information Retrieval*

vol. 11, W. Wu, H. Xiong, and S. Shekhar, Eds. Singapore: Kluwer Academic Publishers, 2003, pp. 105-133.

[5] L. Muflikhah and B. Baharudin, "Clustering of E-Document Using Fuzzy Cluster Approach," in *2nd International Conference on Science & Technology, Application in Industry & Education*, UITM, Pulau Pinang, Malaysia, 2008, pp. 2416-2422.

[6] L. Muflikhah and B. Baharudin, "Optimize Fuzzy Cluster of E-Document using Validity Index," in *International Graduate Conference on Engineering & Science*, UTM Skudai, Johor, Malaysia, 2008, pp. 685-692.

[7] A. K. Jain, M. N. Murty, and P. J. Flynn, "Data Clustering: Review," *ACM Computing Surveys*, 1999.

[8] A. K. Jain and R. C. Dubes, *Algorithms for Clustering Data*. Upper Saddle River, NJ: Prentice Hall, 1988.

[9] Ian H. Witten and Eibe Frank, "Data Mining: Practical Machine Learning Tools and Techniques," in *Data Management Systems*, 2nd ed, J. Gray, Ed. San Fransisco,CA: Morgan Kaufmann, 2005.

[10] J. M. Barnard and G. M. Downs, "Clustering Methods and Their Uses in Computational Chemistry," *Reviews in Computational Chemistry*, vol. 18, pp. 1-40, 2002.

[11] M. Makrechi and M. Shokri, "Document Categorization using Fuzzy Clustering," in *Machine learning course presentation*: Department of SDE, University of Waterloo, 2002.

[12] C. J. v. Rijsbergen, *Information Retrieval*, second ed. Buttersworth, London, 1989.

[13] G. Kowalski, *Information Retrieval Systems*: Kluwer Academic Publisher, 1997.

[14] D. R. Cutting, D. R. Karger, J. O. Pedersenm, and J. W.Tukey, "A Cluster-based Approach to Browsing Large Document Collections," in *In Proceedings of the Fifteenth Annual International ACM SIGIR Conference*, 1992, pp. 318-329.

[15] O. Zemir, O. E. O. Madani, and R. M. Karp, "Fast and Intuitive Clustering of Web Documents," in *Knowledge Discovery and Data Mining*, Uthurusamy, R.Menlo Park, CA, USA, 1997, pp. 287-290.

[16] R. C. Dubes and A. K.Jain, "Algorithms for Clustering Data," Prentice Hall, 1988.

[17] B. Larsen and C. Aone, "Fast and Effective Text Mining Using Linear-time Document Clustering," in *KDD-99*, San Diego, California, 1999, pp. 16-22.

[18] D. Koller and M. Sahami, "Hierarchically Classifying Documents Using Very Few Words," in *Proceedings of the 14th International Conference on Machine Learning ECML98*, 1998.

[19] Richard C. Dubes and Anil K.Jain, "Algorithms for Clustering Data," Prentice Hall, 1988.

[20] Y. El-Sonbaty and M. A. Ismail, "Fuzzy Clustering for Symbol Data," *IEEE Transactions on Fuzzy Systems*, vol. 6, pp. 143-175, 1998.

[21] M. E. S. M. Rodrigues and L. Sacks, "A Scalable Hierarchical Fuzzy Clustering Algorithm for Text Mining," in *The 5th International Conference on Recent Advances in Soft Cpmputing*, 2004.

[22] H. Chi-Hyon Oh, K and Ichihashi,H., "Fuzzy Clustering for Categorical Multivariate Data," in *IFSA World Congress and 20th NAFIPS International Conference, 2001. Joint 9th*, Vancouver, USA, 2001, pp. 2154 - 2159.

[23] Hichem Frigui and Olfa Nasraoui, "Simultaneous Clustering and Dynamic Keyword Weighting for Text Documents," in *Survey of Text Mining*, M. W. Berry, Ed., 2002.

[24] I.S.Dhillon and D.S.Modha, "Concept Decompositions for Large Sparse Text Data using Clustering," *Machine Learning*, vol. 42 pp. 143-175, 2001.

[25] K. Kummamuru, A. Dhawale, and R. Krishnapuram, "Fuzzy Co-clustering of Documents and Keywords," in *The 12th IEEE International Conference on Fuzzy Systems*. vol. 2 St. Louis MO, USA, 2003, pp. 772-777.

[26] P. N. R.Velardi and A. Gangemi, "Ontology Learning and Its Application to Automated Terminology Translation," *IEEE Intelligent Systems,* vol. 18, pp. 22-31, 2003.

[27] V. Sugumaran and V. C. Storey, "Ontologies for Conceptual Modeling: Their Creation, Use and Management," *International Journal of Data and Knowledge Engineering,* vol. 42, pp. 251-271, 2002.

[28] V. W. Soo and C. Y. Lin, "Ontology-based Information Retrieval in A Multi-Agent System for Digital Library," in *Proceedings of the Sixth Conference on Artificial Intelligence and Applications*, Taiwan, 2001, pp. 241-246.

[29] D. H. Wiyantoro and J.Yen, "A Fuzzy Ontology-based Abstract Search Engine and Its User Studies," in *Proceedings of the 10th IEEE International Conference on Fuzzy Systems,* Melbourne, Australia, 2001, pp. 1291-1294.

[30] A. M. A.Hotho, S.Staab, "Ontology-based Text Clustering," in *Proceedings of the IJCAI-2001 Workshop Text Learning: Beyond Supervision*, Seattle, USA, 2001, pp. 30-37.

[31] J. Sedding and D. Kazakov, "WordNet-based Text Document Clustering," in *In Proc. of the Third Workshop on Robust Methods in Analysis of Natural Language Data (ROMAND),* Geneva, 2004, pp. 104-113.

[32] Thomas de Simone and Dimitar Kazakov, "Using WordNet Similarity and Antonymy Relations to Aid Document Retrieval," in *Recent Advance in Natural Language Processing (RANLP)* Borovets, Bulgaria, 2005.

[33] S. W. a. P. S. Chihli Hung, "Hybrid Neural Document Clustering Using Guided Self-Organization and WordNet," *IEEE Intelligent Systems*, vol. 19, pp. 68-77, 2004.

[34] L. Jing, L. Zhou, M.Ng, and J. Huang, "Otology-based Distance Measure for Text Clustering," in *SIAM Text Mining 2006 Workshop*, 2006.

[35] G. Chowdhury, *Introduction to Modern Information Retrieval*: Library Association Publishing London, 2001.

[36] K. Aberer, "EPFL-SSC." vol. 4, L. d. s. d. i. repartis, Ed., 2003, pp. 36-50.

[37] S. Deerwester, "Indexing by latent semantic analysis," *Journal of American Society for Information Science and Technology,* vol. 41, pp. 391-407, 1990.

[38] J. Gao and J. Zhang, "Clustered SVD strategies in latent semantic indexing," *Information Processing & Management,* vol. 41, pp. 1051-1063, 2005.

[39] S. T. Dumais, "LSI meets TREC: A status report," in *The First Text Retrieval Conference*, Gaithersburg, MD, 1993, pp. 137-152.

[40] D. A. Grossman and O. Frieder, *Information Retrieval: Algorithms and Heuristics*: Springer, 2004.

[41] L. A. Zadeh, *Fuzzy Sets* vol. 8, 1965.

[42] F. Hopper, R. Kruse, and T. Runkler, *Fuzzy Cluster Analysis*: John Wiley and Sons, 1999.

[43] J. C. Bezdek, *Pattern Recognition with Fuzzy Objective Function Algorithm*: Plenum Press, 1981.

[44] R. J. Hathaway and J. C. Bezdek, "Recent convergence results for the fuzzy c-means clustering algorithms," *Journal of Classification*, vol. 5, pp. 237-247, 1988.

[45] R. Hathaway, J. Bezdek, and W. Tucker, "An Improved Convergence Theory for the Fuzzy ISODATA Clustering Algorithms," *Analysis of Fuzzy Information*, vol. 3, pp. 123 - 132, 1987.

[46] M. Feher and J. M. Schmidt, "Fuzzy Clustering as a Means of Selecting Representative Conformers and Molecular Alignment," *Journal of Chemical Information and Computer Science*, vol. 43, pp. 810-818, 2003.

[47] H. I. Sadaaki Miyamoto, Katsuhiro Honda, *Algorithm for Fuzzy Clustering* vol. 229. Osaka, Japan: Scientific Publishing Services Pvt. Ltd., Chennai, India, 2008.

[48] A. H. David, "Stemming algorithms: A case study for detailed evaluation," *Journal of the American Society for Information Science*, vol. 47, pp. 70-84, 1996.

[49] T. Korenius, J. Laurikkala, K. Jarvelin, and M. Johula, "Stemming and lemmatization in the clusterng of Finish text documents," in *The ACN-CIKM*, New York, 2004, pp. 625-633.

[50] M. Fuller and J. Zobel, "Conflation-based Comparison of Stemming Algorithms," in *The Third Australian Document Computing Symposium*, Sydney, Australia, 1998.

[51] M. F. Porter, "Lovins revisited," *Charting a New Course: Natural Language Processing and Information Retrieval*, pp. 39-68, 2005.

[52] P. Willett, "The Porter stemming algorithmL then and now," in *Program: electronic library and information system*. vol. 40 (3) Sheffield & York, UK: University of Leeds, 2006, pp. 219-223.

[53] M. Porter, "The Porter Stemming Algorithm," 2006.

[54] G. Salton, *Introduction to Modern Information Retrieval*: McGraw-Hill, 1983.

[55] S. Robertson, *Understanding Inverse Document Frequency: On theoretical Argument for IDF*. 7 JJ Thomson Avenue, Cambridge CB3 OFB, UK, 2004.: Microsoft Research, 2004.

[56] C. D. Manning, P. Raghavan, and H. Schutze, *Introduction to Information Retrieval*. Cambridge: Cambridge University Press, 2008.

[57] T. William-Chandra and C. Lihui, "A heuristic-based fuzzy co-clustering algorithm for categorization of high-dimensional data," *Fuzzy Sets Syst.*, vol. 159, pp. 371-389, 2008.

[58] D. Cheng, R. Kannan, S. Vempala, and G. Wang, "A divide-and-merge methodology for clustering," *ACM Transactions on Database System*, vol. 31, pp. 1499-1525, 2006.

[59] J. Kogan, *Introduction to Clustering Large and High-Dimensional Data*. Cambridge: Cambridge University Press, 2007.

[60] C.-H. Oh, K. Honda, and H. Ichihashi, "Fuzzy clustering for categorical multivariate data," in *IFSA World Congress and 20th NAFIPS International Conference, 2001. Joint 9th*, 2001, pp. 2154-2159 vol.4.

[61] H. Frigui and O. Nasraoui, "Simultaneous Clustering and Dynamic Keyword Weighting for Text Documents," in *Survey of Text Mining*, M. W. Berry, Ed., 2002.

LIST OF PUBLICATIONS

Year	Title	Publication In
2008	Clustering of E-Documents Using Fuzzy Cluster Approach Co Author : Baharum Baharudin	Proceeding of 2nd International Conference on Science & Technology, Application in Industry & Education (2008), pp. 2416-2422, UITM, Pulau Pinang, Malaysia (12-13 December 2008)
2008	Optimize Fuzzy Cluster of E-Documents using Validity Index Co Author : Baharum Baharudin	International Graduate Conference on Engineering & Science, pp. 685-692 , UTM Skudai, Johor, Malaysia (23-24 December 2008)
2009	Improvement of Fuzzy Clustering Method for E-Document Clustering Co Author : Baharum Baharudin	Proceeding of National Postgraduate Conference on Engineering, Science and Technology (NPC 2009), Universiti Teknologi PETRONAS, Malaysia (25-26 March 2009)
2009	High Performance in Minimizing of Term-Document Matrix Representation for Document Clustering Co Author : Baharum Baharudin	Proceeding of Conference & Exhibition on Innovative Technologies in Intelligent System & Industrial Applications (CITISIA 2009), Monash University, Malaysia (25-26 July 2009) – *indexed by IEEE*
2009	Document Clustering Using Concept Space and Cosine Similarity Measuremen Co Author : Baharum Baharudin	Proceeding of the 2009 International Conference on Computer Technology and Development (ICCTD), Kota Kinabalu, Malaysia (13-15 November 2009) – *indexed by IEEE*
2010	Improving Performance using Latent Semantic Indexing for Document Categorization Co Author : Baharum Baharudin	IAENG International Journal of Computer Science (25 January, 2010)– *will be indexed by Scopus*

APPENDIX A

Stop Word List

A	become	early	gives	itself
about	becomes	either	go	j
above	been	end	going	just
across	before	ended	good	k
after	began	ending	goods	keep
again	behind	ends	got	keeps
against	being	enough	great	kind
all	beings	even	greater	knew
almost	best	evenly	greatest	know
alone	better	ever	group	known
along	between	every	grouped	knows
already	big	everybody	grouping	l
also	both	everyone	groups	large
although	but	everything	h	largely
always	by	everywhere	had	last
among	c	f	has	knows
an	came	face	have	l
and	can	faces	having	large
another	cannot	fact	he	largely
any	case	facts	her	last
anybody	cases	far	here	later
anyone	certain	felt	herself	latest
anything	certainly	few	high	least
anywhere	clear	find	high	less
are	clearly	finds	high	let
area	come	first	higher	lets
areas	could	for	highest	like
around	d	four	him	likely
as	did	from	himself	long
ask	differ	full	his	longer
asked	different	fully	how	longest
asking	differently	further	however	m
asks	do	furthered	i	made
at	does	furthering	if	make
away	done	furthers	important	making
b	down	g	in	man
back	down	gave	interest	many
backed	downed	general	interested	may
backing	downing	generally	interesting	me
backs	downs	get	interests	member
be	during	gets	into	members
became	e	give	is	men
because	each	given	it	might

more
most
mostly
mr
mrs
much
must
my
myself
n
necessary
need
needed
needing
needs
never
new
newer
newest
next
no
nobody
non
noone
not
nothing
now
nowhere
number
numbers
o
of
off
often
old
older
oldest
on
once
one
only
open
opened
opening
opens
or
order
ordered
ordering
orders
other
others
our
out
over
p
part
parted
parting
parts
per
perhaps
place
places
point
pointed
pointing
points
possible
present
presented
presenting
presents
problem
problems
put
puts
q
quite
r
rather
really
right
right
room
rooms
s
said
same
saw
say
says
second
seconds
see
seem
seemed
seeming
seems
sees
several
shall
she
should
show
showed
showing
shows
side
sides
since
small
smaller
smallest
so
some
somebody
someone
something
somewhere
state
states
still
still
such
sure
t
take
taken
than
that
the
their
them
then
there
therefore
these
they
thing
things
think
thinks
this
those
though
thought
thoughts
three
through
thus
to
today
together
too
took
toward
turn
turned
turning
turns
two
u
use
used
uses
v
very
w
want
wanted
wanting
wants
was
way
ways
we
well
wells
went
were
what
when
where
whether
which
while
who
whole
whose
why
will
with
within
without
whole
whose
why
will
with
within
without

APPENDIX B

Document Form

Data set: 20News Group; Category: politics.mideast

Xref: cantaloupe.srv.cs.cmu.edu soc.culture.turkish:32574 talk.politics.soviet:22956
talk.politics.mideast:75369 soc.culture.greek:21385
Path: cantaloupe.srv.cs.cmu.edu!das-news.harvard.edu!husc-
news.harvard.edu!hsdndev!wupost!uunet!vuse.vanderbilt.edu!senel
Newsgroups: soc.culture.turkish,talk.politics.soviet,talk.politics.mideast,soc.culture.greek
Subject: Re: If You Feed Armenians Dirt -- You Will Bite Dust!
Message-ID: <senel.2@vuse.vanderbilt.edu>
From: senel@vuse.vanderbilt.edu (Hakan)
Date: 5 Apr 93 20:45:13 GMT
Sender: news@vuse.vanderbilt.edu
References: <1993Apr5.194120.7010@urartu.sdpa.org>
Organization: Vanderbilt University
Summary: Armenians correcting the geo-political record.
Nntp-Posting-Host: snarl02
Lines: 18

In article <1993Apr5.194120.7010@urartu.sdpa.org> dbd@urartu.sdpa.org (David Davidian)
writes:

>In article <1993Apr5.064028.24746@kth.se> hilmi-er@dsv.su.se (Hilmi Eren)
>writes:

>David Davidian says: Armenians have nothing to lose! They lack food, fuel, and
>warmth. If you fascists in Turkey want to show your teeth, good for you! Turkey
>has everything to lose! You can yell and scream like barking dogs along the

Davidian, who are fascists? Armenians in Azerbaijan are killing Azeri
people, invading Azeri soil and they are not fascists, because they
lack food ha? Strange explanation. There is no excuse for this situation.

Herkesi fasist diye damgala sonra, kendileri fasistligin alasini yapinca,
"ac kaldilar da, yiyecekleri yok amcasi, bu seferlik affedin" de. Yurrruuu,
yuru de plaka numarani alalim.....

Hakan

Data set: Classic3; Category: Information retrieval

.I 1

18 Editions of the Dewey Decimal Classifications
.A
Comaromi, J.P.
.W
  The present study is a history of the DEWEY Decimal Classification. The first edition of the DDC was published in 1876, the eighteenth edition in 1971, and future editions will continue to appear as needed. In spite of the DDC's long and healthy life, however, its full story has never been told. There have been biographies of Dewey that briefly describe his system, but this is the first attempt to provide a detailed history of the work that more than any other has spurred the growth of librarianship in this country and abroad.
.X
1 2 1
1 2 1
1 2 1
1 2 1
1 2 1
556 2 1
92 2 1
262 2 1
1004 2 1
1024 2 1
.I 2
</TEXT>
Use Made of Technical Libraries
.A
Slater, M.
.W
This report is an analysis of 6300 acts of use in 104 technical libraries in the United Kingdom. Library use is only one aspect of the wider pattern of information use. Information transfer in libraries is restricted to the use of documents. It takes no account of documents used outside the library, still less of information transferred orally from person to person. The library acts as a channel in only a proportion of the situations in which information is transferred.

YahooK1 (doc000011) – Category: Health

Wednesday October 8 6:39 PM EDT

House Passes FDA Overhaul Bill

WASHINGTON (Reuters) -- The U.S. House of Representatives passed the once controversial bill to streamline and overhaul operations at the Food and Drug Administration (FDA) by voice vote Tuesday.

The bill passed once Republicans and Democrats worked out compromises on a wide array of issues of concern to both the agency and the drug and device makers it regulates. The bill now goes to conference with the Senate, which passed its version of the measure September 24th.

Congress is under pressure to complete action on the bill, because it includes a five-year reauthorization of the Prescription Drug User Fee Act, which technically expired on October 1.

Citing the importance of the renewal of the popular program that has helped cut review times for new drugs almost in half, the Clinton administration issued a statement Tuesday morning saying it "...has no objection..." to passage of the bill "...at this time." But the statement did go on to say that the administration "...continues to have major concerns with the bill," including a contention that the provision allowing "third-party" review of medical devices is "too broad," and could permit even potentially dangerous devices to be examined by non-FDA reviewers.

Republicans and Democrats in the House, however, heaped praise on the measure. "We have given the FDA the tools it needs to alleviate the suffering of American patients," said Rep. Michael Bilirakis (R-Florida), chairman of the subcommittee that oversees FDA. Added Rep. John Dingell (D-Michigan), "This bill serves the interests of the consuming public."

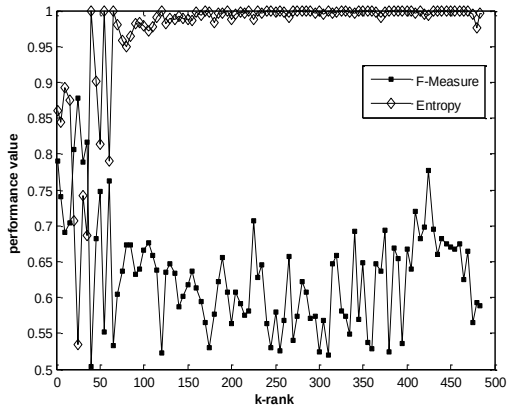
The bipartisan show of support was made possible by two last-minute compromises worked out since the House Commerce Committee approved the measure in September. One would permit FDA to order medical device manufacturers to extend post-market surveillance by up to three years, and even longer under certain conditions. The other would refine a compromise reached earlier regarding the agency's ability to look beyond manufacturers' stated use of a device in determining safety and efficacy.

Under the compromise, a device could continue to be marketed for the purpose stated on the label, even if the FDA believes that it will be used for another purpose. This can occur as long as the label states the alternative purpose as an explicit contraindication while the safety and efficacy of that alternative purpose is evaluated.

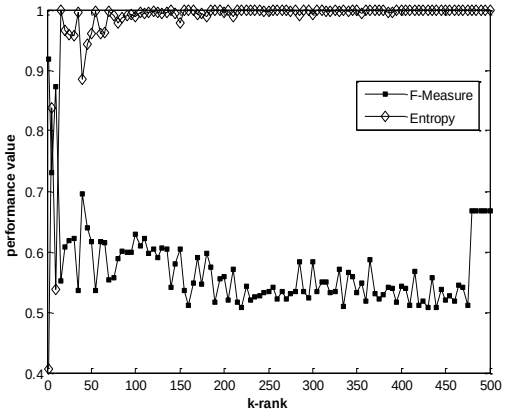
APPENDIX C

Performance of Document Clustering with Various *k*-Rank

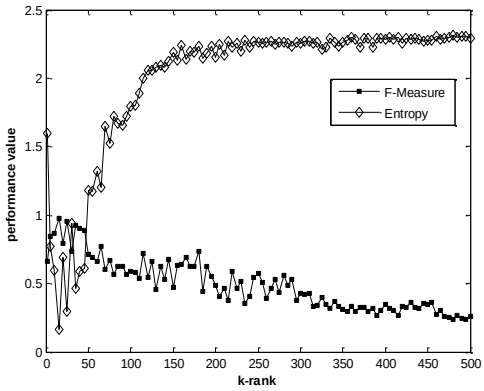
Method: SVD; data set: binary2



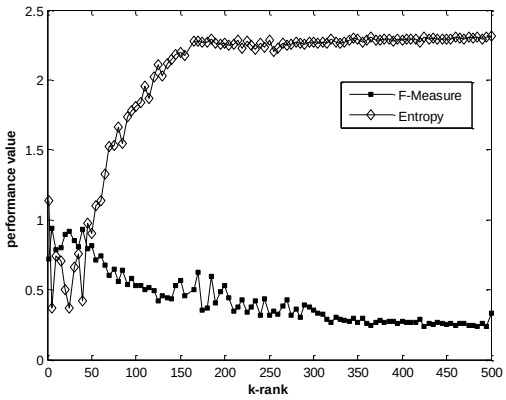
Method: PCA; data set binary2



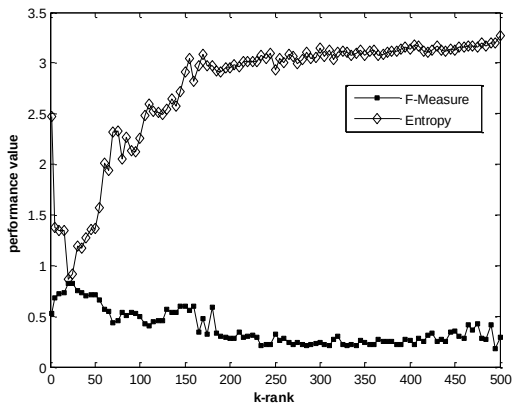
Method: SVD; data set: Multi5



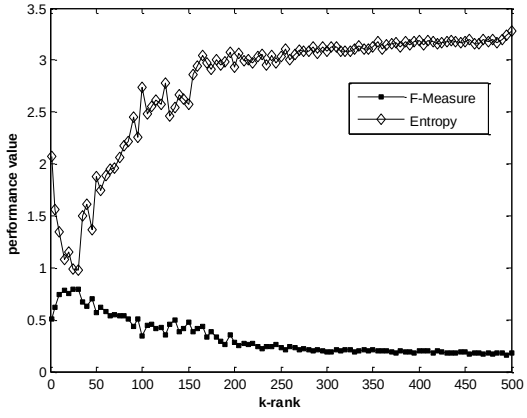
Method: PCA; data set: Multi5



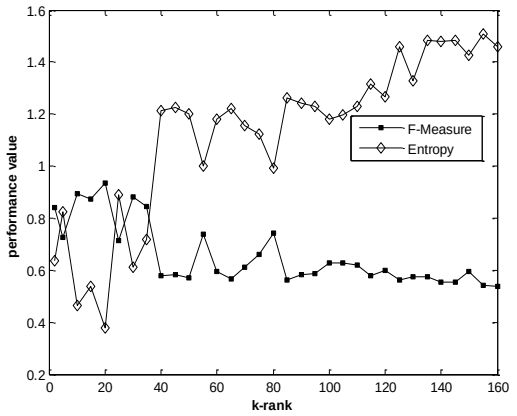
Method: SVD; data set Multi10



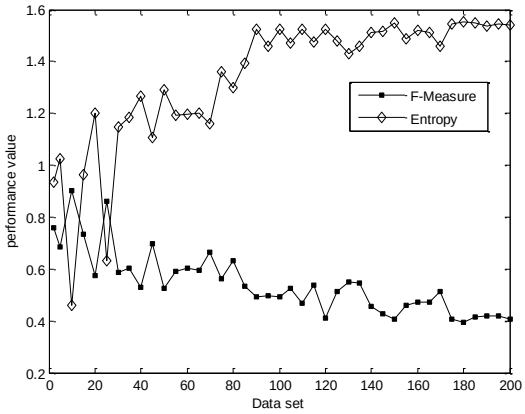
Method: PCA; data set: Multi10



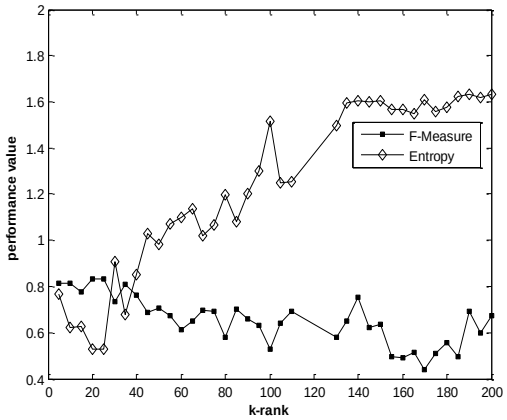
Method: SVD; data set Classic3



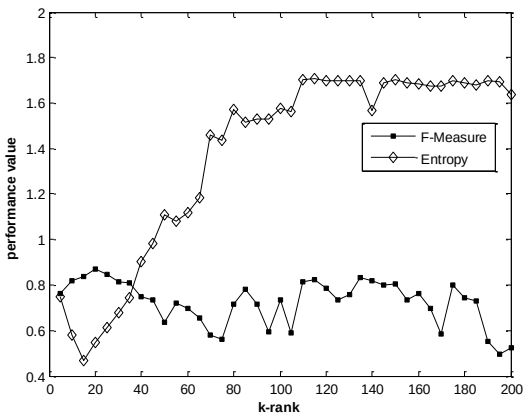
Method: PCA; data set Classic3



Method: SVD; data set YahooK1

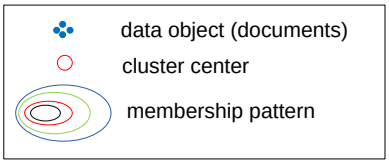


Method: PCA; data set YahooK1

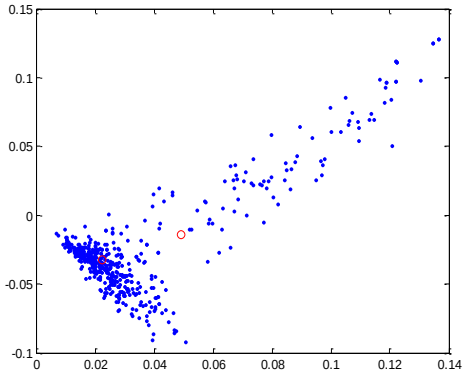


APPENDIX D

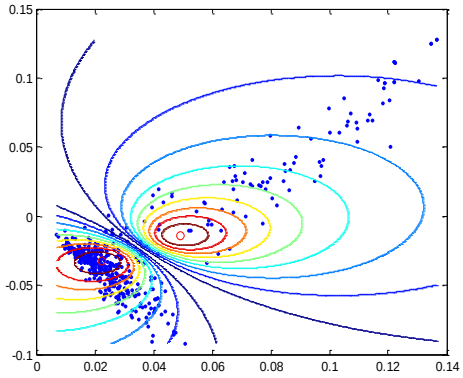
Result Illustration of Fuzzy c-Means Clustering Algorithm



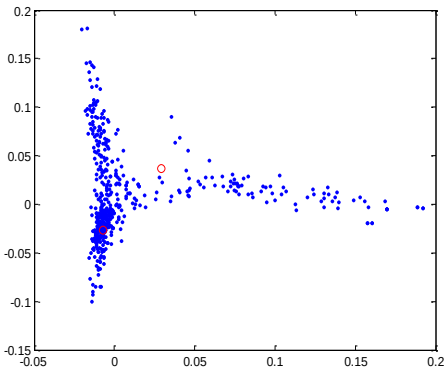
Data set binary2; method: SVD



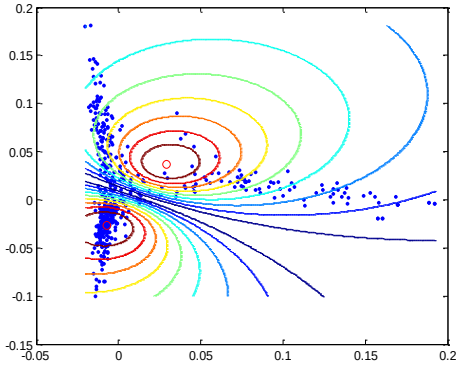
Data set binary2; method: SVD



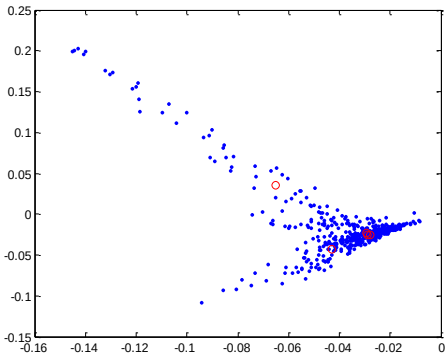
Data set binary2; method:PCA



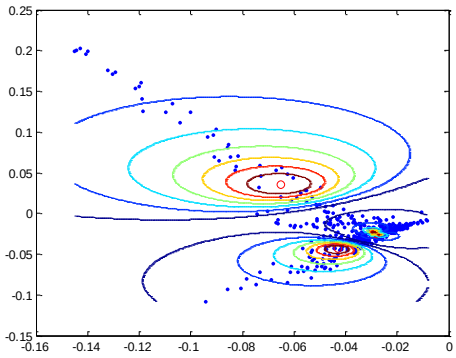
Data set binary2; method:PCA;



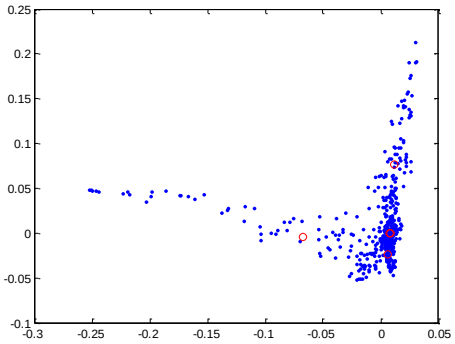
Data set multi5; method:SVD



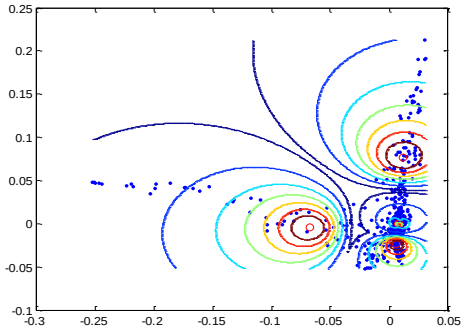
Data set multi5; method SVD



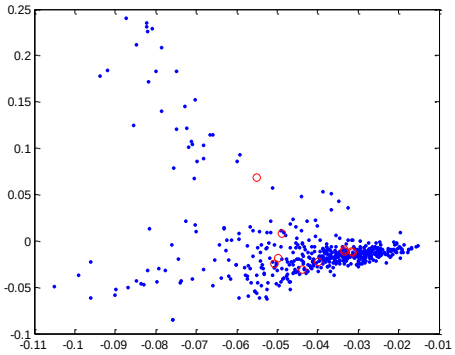
Data set multi5; method:PCA;



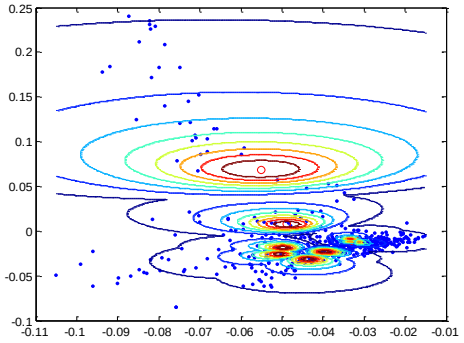
Data set multi5; method:PCA;



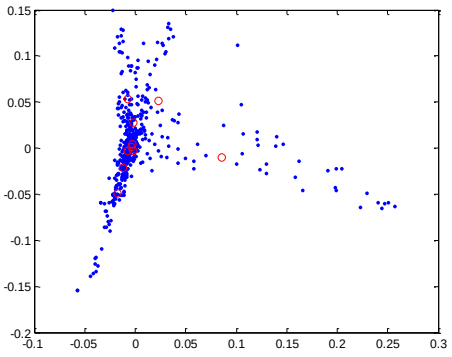
Data set -multi10; method: SVD;



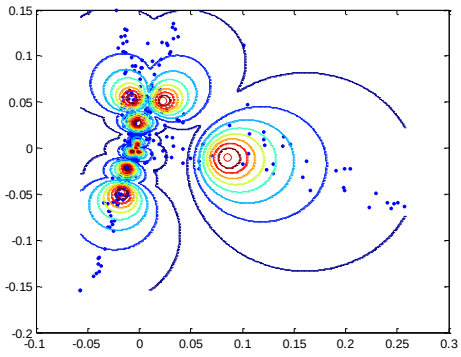
Data set multi10; method: SVD;

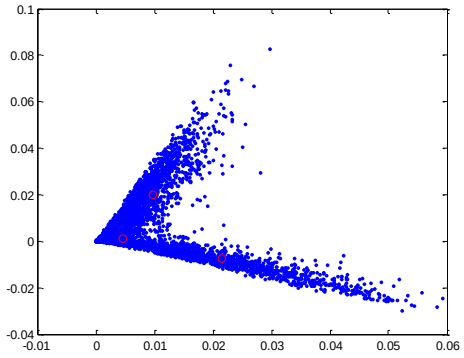
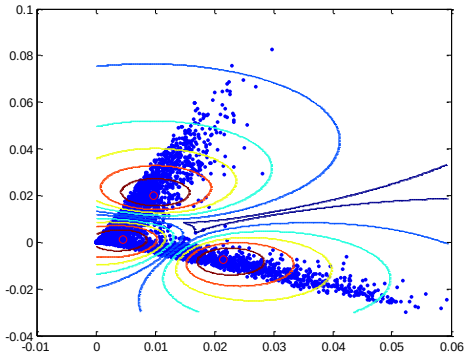
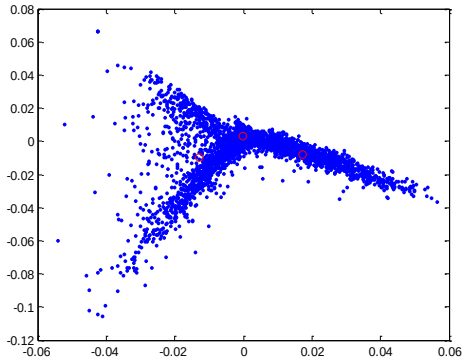
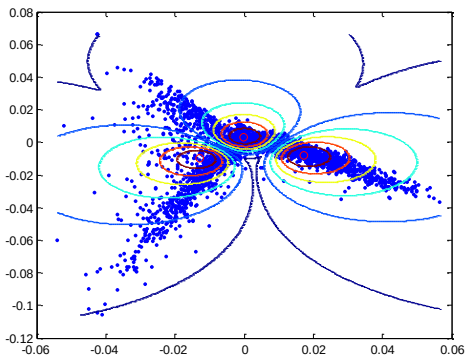
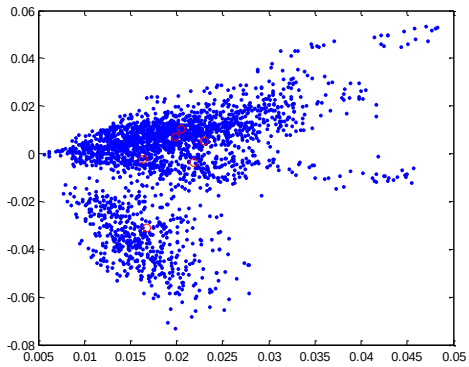
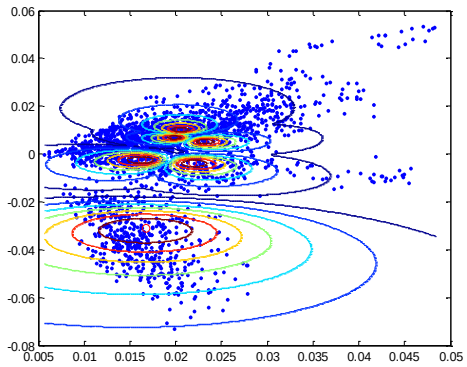
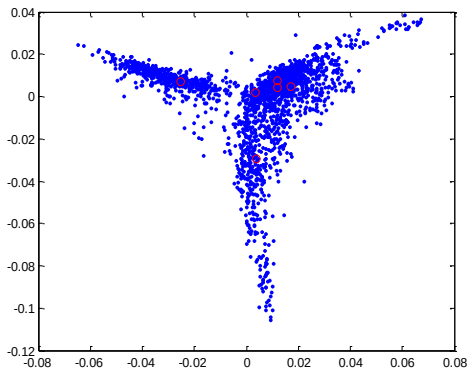
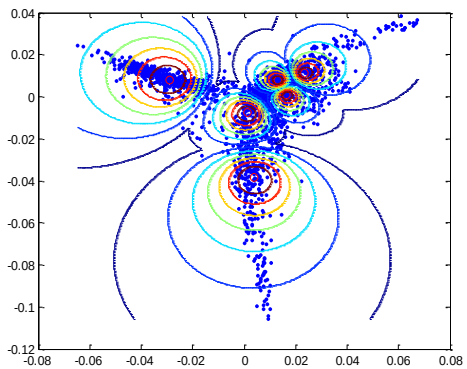


Data set multi10; method:PCA;



Data set multi10; method:PCA;



Data set Classic3; method: SVD	Data set Classic3; method: SVD
	
Data set Classic3; method: PCA	Data set Classic3; method: PCA
	
Data set YahooK1; method: SVD	Data set YahooK1; method: SVD
	
Data set YahooK1; method: PCA	Data set YahooK1; method: PCA
	

APPENDIX E

Document Clustering Result

E.1. SVD

Data set 'Binary2' (▭ = misclassified document)

Cluster No.	Document Number														
cluster #1	1	2	3	4	5	6	7	8	9	10	14	15	16	17	19
	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34
	35	36	38	39	41	42	43	44	46	47	49	51	52	53	54
	55	56	57	58	59	60	61	62	63	64	65	66	67	68	69
	70	71	72	73	74	75	76	77	78	79	80	81	82	83	84
	85	86	87	88	89	90	92	93	94	95	96	97	99	100	101
	102	103	104	105	106	107	108	109	110	111	112	113	114	115	116
	117	118	119	120	121	122	124	125	126	127	128	129	130	131	132
	133	134	135	136	137	138	139	140	141	142	143	144	145	146	148
	149	150	151	152	153	154	156	157	158	159	160	161	162	163	164
	165	166	167	168	169	170	171	172	173	174	175	176	177	178	179
	180	183	184	185	186	187	188	189	190	191	192	193	194	195	196
	197	198	199	200	201	202	203	204	205	206	207	208	209	210	211
	212	213	214	216	217	218	219	220	221	222	223	224	225	226	227
	229	230	231	232	233	234	235	236	237	240	241	242	243	244	245
	246	247	248	249	250	258	260	278	297	298	313	378	380	388	397
	402	403	404	419	434	440	442	474	482	488					
cluster #2	11	12	13	18	37	40	45	48	50	91	98	123	147	155	181
	182	215	228	238	239	251	252	253	254	255	256	257	259	261	262
	263	264	265	266	267	268	269	270	271	272	273	274	275	276	277
	279	280	281	282	283	284	285	286	287	288	289	290	291	292	293
	294	295	296	299	300	301	302	303	304	305	306	307	308	309	310
	311	312	314	315	316	317	318	319	320	321	322	323	324	325	326
	327	328	329	330	331	332	333	334	335	336	337	338	339	340	341
	342	343	344	345	346	347	348	349	350	351	352	353	354	355	356
	357	358	359	360	361	362	363	364	365	366	367	368	369	370	371
	372	373	374	375	376	377	379	381	382	383	384	385	386	387	389
	390	391	392	393	394	395	396	398	399	400	401	405	406	407	408
	409	410	411	412	413	414	415	416	417	418	420	421	422	423	424
	425	426	427	428	429	430	431	432	433	435	436	437	438	439	441
	443	444	445	446	447	448	449	450	451	452	453	454	455	456	457
	458	459	460	461	462	463	464	465	466	467	468	469	470	471	472
	473	475	476	477	478	479	480	481	483	484	485	486	487	489	490
	491	492	493	494	495	496	497	498	499	500					

Data set ‘Multi5’ (= misclassified document)

Cluster No.	Document Number														
cluster #1	201	202	203	204	205	206	207	208	209	210	211	212	213	214	215
	216	217	218	219	220	221	222	223	224	225	226	227	228	229	230
	231	232	233	234	235	236	237	238	239	240	241	242	243	244	245
	246	247	248	249	251	252	253	254	255	256	257	258	259	260	261
	262	263	264	265	266	267	268	269	270	271	272	273	274	275	277
	278	279	280	281	282	283	284	285	286	287	288	289	290	292	294
	295	296	298	299	300										
cluster #2	101	102	103	104	105	106	107	108	109	110	111	112	113	114	115
	116	117	118	119	120	121	122	123	124	125	126	127	128	129	130
	131	132	133	134	135	136	137	138	139	140	141	142	143	144	145
	146	147	148	149	150	151	152	153	154	155	156	157	158	159	160
	161	162	163	164	165	166	167	168	169	170	172	173	174	175	176
	177	178	179	181	182	183	184	186	187	188	189	190	191	192	193
	194	195	196	197	198	199	200	276	293						
cluster #3	291	301	302	303	304	305	306	307	308	309	310	311	312	313	314
	315	316	317	318	319	320	321	322	323	324	325	326	327	328	329
	330	331	332	333	334	335	336	337	338	339	340	341	342	343	344
	345	346	347	348	349	350	351	352	353	354	355	357	358	359	360
	361	362	363	364	365	366	368	369	371	372	373	374	375	376	377
	378	379	380	381	382	383	384	385	386	387	388	389	390	391	392
	393	394	395	396	397	398	399	400							
cluster #4	180	401	402	403	404	405	406	407	408	409	410	411	412	413	414
	415	416	417	418	419	420	421	422	423	424	425	426	427	428	429
	430	431	432	433	434	435	436	437	438	439	440	441	442	443	444
	445	446	447	448	449	450	451	452	453	454	455	456	457	458	459
	460	461	462	463	464	465	466	467	468	469	470	471	472	473	474
	475	476	477	478	479	480	481	482	483	484	485	486	487	488	489
	490	491	492	493	494	495	496	497	498	499	500				
cluster #5	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45
	46	47	48	49	50	51	52	53	54	55	56	57	58	59	60
	61	62	63	64	65	66	67	68	69	70	71	72	73	74	75
	76	77	78	79	80	81	82	83	84	85	86	87	88	89	90
	91	92	93	94	95	96	97	98	99	100	171	185	250	297	356
	367	370													

Data set ‘Multi10’ (= misclassified document)

Cluster No.	Document Number														
cluster #1	232	233	248	251	252	253	254	255	256	257	258	259	260	261	262
	263	264	265	268	269	270	271	272	273	274	275	276	277	278	279
	280	281	282	283	284	285	286	287	288	289	290	291	292	293	294
	295	296	297	298	299	300	301	308	343	352					
cluster #2	29	52	57	83	101	102	103	104	105	106	108	109	111	112	113
	115	116	117	118	119	120	121	122	123	124	125	126	127	128	129
	131	132	133	134	135	136	137	138	139	140	141	142	143	144	146
	147	148	149	150	205	208	227	235	307	353	375	420	444		
cluster #3	145	201	202	203	204	206	209	210	211	212	213	214	215	216	217
	218	219	220	221	222	223	224	225	226	228	229	230	231	234	236
	237	238	240	241	243	244	245	246	247	249	250	355			
cluster #4	23	24	239	303	310	311	313	315	318	322	325	326	329	330	331
	332	333	334	335	336	337	339	340	341	342	344	349	357	367	369
	371	372	400	421	425	431	439	440	442	443	445	450			
cluster #5	18	32	93	151	152	153	154	155	156	157	158	159	160	161	162
	163	164	165	166	167	168	169	170	171	172	173	174	175	176	177
	178	179	180	181	182	183	185	186	187	188	189	190	191	192	193
	194	195	196	197	198	199	200	207	242	302	304	305	306	309	312
	314	358	377												
cluster #6	5	114	130	316	317	319	327	338	347	348	351	354	356	359	360
	363	364	365	366	368	370	373	374	376	378	379	380	381	382	383
	384	385	386	387	388	389	390	391	392	393	394	395	396	397	398
	399	435													
cluster #7	39	88	107	110	321	423	446	451	452	453	454	455	456	457	458
	459	460	461	462	463	464	465	466	467	468	469	470	471	472	473
	474	475	476	477	478	479	480	481	482	483	484	485	486	487	488
	489	490	491	492	493	494	495	496	497	498	499	500			
cluster #8	51	53	54	55	56	58	59	60	61	62	63	64	65	66	67
	68	69	70	71	72	73	74	75	76	77	78	79	80	81	82
	84	85	86	87	89	90	91	92	94	95	96	97	98	99	100
	184	266	323	324	328	345	346	350	361	362					
cluster #9	35	267	320	401	402	403	404	405	406	407	408	409	410	411	412
	413	414	415	416	417	418	419	422	424	426	427	428	429	430	432
	433	434	436	437	438	441	447	448	449						
cluster #10	1	2	3	4	6	7	8	9	10	11	12	13	14	15	16
	17	19	20	21	22	25	26	27	28	30	31	33	34	36	37
	38	40	41	42	43	44	45	46	47	48	49	50			

Data set ‘Classic3’(= misclassified document)

Cluster No.	Document Number														
Cluster # 1	233	306	335	347	367	423	439	462	492	534	547	663	716	789	1010
	1037	1082	1084	1115	1195	1214	1228	1239	1310	1323	1372	1384	1385	1397	1415
	1447	1448	1449	1450	1451	1452	1453	1454	1455	1456	1457	1458	1459	1460	1461
	1462	1463	1464	1465	1466	1467	1468	1469	1470	1471	1472	1473	1474	1475	1476
	1477	1478	1479	1480	1481	1482	1483	1484	1485	1486	1487	1488	1489	1490	1491
	1493	1494	1495	1496	1497	1498	1499	1500	1501	1502	1503	1504	1505	1506	1507
	1508	1509	1510	1511	1512	1513	1514	1515	1516	1517	1518	1519	1520	1522	1523
	1524	1525	1526	1527	1528	1530	1531	1532	1533	1534	1535	1536	1537	1538	1539
	1540	1541	1542	1543	1544	1545	1546	1547	1548	1549	1550	1551	1552	1553	1554
	1555	1556	1557	1558	1559	1560	1561	1562	1563	1564	1565	1566	1567	1568	1569
	1570	1571	1572	1573	1574	1575	1577	1578	1579	1580	1581	1582	1583	1584	1585
	1586	1587	1588	1590	1591	1592	1593	1594	1595	1596	1597	1598	1599	1600	1601
	1603	1604	1605	1606	1607	1608	1609	1610	1611	1612	1613	1614	1615	1616	1617
	1618	1619	1620	1621	1622	1623	1624	1625	1626	1627	1628	1629	1631	1632	1633
	1634	1635	1636	1637	1638	1639	1640	1641	1642	1643	1644	1645	1646	1647	1648
	1649	1650	1651	1652	1653	1654	1655	1656	1657	1658	1659	1660	1661	1662	1663
	1664	1665	1666	1667	1668	1669	1670	1671	1672	1673	1674	1675	1676	1677	1678
	1679	1680	1681	1682	1684	1685	1686	1687	1688	1689	1690	1691	1692	1693	1694
	1695	1696	1697	1698	1699	1700	1701	1702	1703	1704	1705	1706	1707	1708	1709
	1710	1711	1712	1713	1714	1715	1716	1717	1718	1719	1720	1721	1722	1723	1724
	1725	1726	1727	1728	1729	1730	1731	1732	1733	1734	1735	1736	1737	1738	1739
	1740	1741	1742	1743	1744	1745	1746	1747	1748	1749	1750	1751	1752	1753	1754
	1755	1756	1757	1758	1759	1760	1761	1762	1763	1764	1765	1766	1767	1768	1769
	1770	1771	1772	1773	1774	1775	1777	1778	1779	1780	1781	1782	1783	1784	1785
	1786	1787	1788	1789	1790	1791	1792	1793	1794	1795	1796	1797	1798	1799	1800
	1801	1802	1803	1804	1805	1806	1807	1808	1809	1810	1811	1812	1815	1816	1817
	1818	1819	1820	1821	1822	1823	1824	1825	1826	1827	1828	1829	1830	1831	1832
	1833	1834	1835	1836	1837	1838	1839	1840	1841	1842	1843	1844	1846	1847	1848
	1849	1850	1851	1852	1853	1854	1855	1856	1857	1858	1859	1860	1861	1862	1863
	1864	1865	1866	1867	1868	1869	1870	1871	1872	1873	1874	1875	1876	1877	1878
	1879	1880	1881	1882	1883	1884	1885	1886	1887	1888	1889	1890	1891	1892	1893
	1894	1895	1896	1897	1898	1899	1900	1901	1902	1903	1904	1905	1906	1907	1908
	1909	1910	1911	1912	1913	1914	1915	1916	1917	1918	1919	1920	1921	1922	1923
	1924	1925	1926	1927	1928	1929	1930	1931	1932	1934	1935	1936	1937	1938	1939
	1940	1941	1942	1943	1944	1945	1946	1947	1948	1949	1950	1952	1953	1954	1955
	1956	1957	1958	1959	1960	1961	1963	1964	1965	1966	1967	1968	1969	1970	1971
	1972	1973	1974	1975	1976	1977	1978	1979	1980	1981	1982	1983	1984	1985	1986
	1987	1989	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002
	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018
	2019	2020	2021	2022	2023	2024	2025	2026	2027	2028	2029	2030	2031	2032	2033
	2034	2035	2036	2037	2038	2039	2040	2041	2042	2044	2045	2046	2047	2048	2049

Cluster No.	Document Number														
	2050	2051	2052	2053	2054	2055	2056	2057	2058	2059	2060	2063	2064	2066	2068
	2069	2070	2071	2072	2073	2074	2075	2076	2077	2078	2079	2080	2081	2082	2083
	2084	2086	2087	2088	2089	2090	2091	2092	2093	2095	2096	2097	2098	2099	2100
	2101	2102	2103	2104	2105	2106	2107	2108	2109	2110	2111	2112	2113	2114	2115
	2116	2117	2118	2119	2120	2121	2122	2123	2124	2125	2126	2127	2128	2129	2130
	2131	2132	2133	2134	2135	2136	2137	2138	2139	2140	2141	2142	2143	2144	2145
	2146	2147	2148	2149	2150	2151	2152	2153	2154	2155	2156	2157	2158	2159	2160
	2162	2164	2165	2166	2167	2168	2169	2171	2172	2173	2174	2175	2176	2177	2178
	2179	2180	2181	2182	2183	2184	2185	2186	2187	2188	2189	2191	2192	2193	2194
	2195	2196	2197	2198	2199	2200	2201	2202	2203	2204	2205	2206	2207	2208	2209
	2210	2211	2212	2213	2214	2215	2216	2217	2218	2219	2220	2221	2222	2223	2224
	2225	2226	2227	2228	2229	2230	2231	2232	2233	2234	2235	2236	2237	2238	2239
	2240	2241	2242	2243	2244	2245	2246	2247	2248	2249	2250	2251	2252	2253	2254
	2255	2256	2257	2258	2259	2260	2261	2262	2263	2264	2265	2266	2267	2268	2269
	2270	2271	2272	2273	2274	2275	2276	2277	2278	2281	2282	2283	2284	2285	2286
	2287	2288	2289	2290	2291	2292	2293	2294	2295	2296	2297	2298	2299	2300	2301
	2302	2303	2304	2305	2306	2307	2308	2309	2310	2311	2313	2314	2315	2316	2317
	2318	2319	2321	2322	2323	2324	2325	2326	2327	2330	2331	2332	2333	2334	2335
	2336	2337	2338	2339	2340	2341	2342	2343	2344	2345	2346	2347	2348	2349	2351
	2352	2353	2354	2355	2357	2359	2360	2361	2362	2363	2364	2365	2366	2367	2368
	2369	2370	2371	2372	2373	2374	2375	2376	2377	2378	2379	2380	2381	2382	2383
	2384	2385	2386	2388	2389	2390	2391	2392	2393	2394	2395	2396	2397	2398	2399
	2400	2401	2402	2404	2405	2406	2407	2408	2409	2410	2411	2412	2414	2415	2416
	2417	2418	2419	2420	2421	2422	2423	2424	2425	2426	2427	2428	2429	2430	2431
	2432	2433	2434	2435	2436	2437	2438	2439	2440	2441	2442	2443	2444	2445	2446
	2447	2448	2449	2450	2451	2452	2453	2454	2455	2456	2457	2458	2459	2460	2461
	2462	2463	2464	2465	2466	2467	2468	2469	2470	2471	2472	2473	2474	2475	2476
	2477	2478	2479	2480	2481	2482	2483	2484	2485	2486	2487	2488	2489	2490	2491
	2492	2493	2494	2495	2496	2497	2498	2499	2500	2501	2502	2503	2504	2505	2506
	2507	2508	2509	2510	2511	2512	2513	2514	2515	2516	2517	2518	2519	2520	2521
	2522	2523	2524	2525	2526	2527	2528	2529	2530	2531	2532	2533	2534	2535	2536
	2537	2538	2539	2540	2542	2543	2544	2548	2549	2550	2551	2552	2553	2554	2555
	2556	2558	2559	2560	2561	2562	2563	2564	2565	2566	2567	2568	2569	2570	2571
	2572	2573	2574	2575	2576	2577	2578	2579	2580	2581	2582	2583	2584	2585	2586
	2587	2588	2589	2590	2591	2592	2593	2594	2595	2597	2598	2599	2600	2601	2602
	2603	2605	2606	2607	2608	2609	2610	2612	2613	2614	2615	2616	2617	2618	2619
	2620	2621	2622	2623	2624	2625	2626	2627	2628	2629	2630	2631	2632	2633	2634
	2635	2636	2637	2638	2639	2641	2642	2643	2644	2645	2646	2647	2648	2649	2650
	2651	2652	2653	2654	2655	2656	2657	2658	2659	2660	2661	2662	2663	2664	2665
	2666	2667	2668	2669	2670	2671	2672	2673	2674	2675	2676	2677	2678	2679	2680
	2681	2682	2683	2684	2685	2686	2687	2688	2689	2690	2691	2692	2693	2694	2695
	2696	2697	2698	2699	2700	2701	2702	2703	2704	2705	2706	2707	2708	2709	2710
	2711	2712	2713	2714	2715	2716	2717	2718	2719	2720	2721	2722	2723	2724	2725

Cluster No.	Document Number														
	2726	2727	2728	2729	2730	2731	2733	2734	2735	2736	2737	2738	2739	2740	2741
	2742	2743	2744	2745	2746	2747	2748	2749	2750	2751	2752	2753	2754	2755	2756
	2757	2758	2759	2760	2761	2762	2763	2764	2765	2766	2767	2768	2769	2770	2771
	2772	2773	2774	2775	2776	2777	2778	2779	2780	2781	2782	2783	2784	2785	2786
	2787	2788	2789	2790	2791	2792	2793	2794	2795	2796	2797	2798	2799	2800	2801
	2802	2803	2804	2805	2806	2807	2808	2809	2810	2811	2812	2813	2814	2815	2816
	2817	2818	2819	2820	2821	2822	2823	2824	2825	2826	2827	2828	2829	2830	2831
	2833	2834	2835	2836	2837	2838	2839	2840	2841	2842	2843	2844	2877	2971	2973
	3075	3198	3225	3262	3264	3450	3459	3460	3462	3529	3552	3557	3604	3668	3808
	3853														
Cluster # 2	30	35	94	99	101	108	118	157	168	227	229	231	256	389	414
	415	416	428	471	556	603	626	665	743	748	798	860	888	891	900
	909	976	991	1025	1027	1032	1041	1042	1046	1059	1061	1065	1078	1089	1113
	1137	1146	1150	1153	1155	1164	1182	1184	1251	1264	1268	1277	1281	1290	1292
	1297	1300	1302	1316	1319	1321	1327	1329	1367	1375	1392	1423	1437	1442	1521
	1529	1589	1602	1683	1776	1814	1845	1933	1951	1962	1988	2003	2061	2062	2065
	2067	2085	2094	2161	2170	2279	2280	2312	2328	2329	2350	2356	2358	2387	2403
	2541	2545	2547	2557	2596	2604	2611	2640	2732	2845	2846	2847	2848	2849	2850
	2851	2852	2853	2854	2855	2856	2857	2858	2859	2860	2861	2862	2863	2864	2865
	2866	2867	2868	2869	2870	2871	2872	2873	2874	2875	2876	2878	2879	2880	2881
	2882	2883	2884	2885	2886	2887	2888	2889	2890	2891	2892	2893	2894	2895	2896
	2897	2898	2899	2900	2901	2902	2903	2904	2905	2906	2907	2908	2909	2910	2911
	2912	2913	2914	2915	2916	2917	2918	2919	2920	2921	2922	2923	2924	2925	2926
	2927	2928	2929	2930	2931	2932	2933	2934	2935	2936	2937	2938	2939	2940	2941
	2942	2943	2944	2945	2946	2947	2948	2949	2950	2951	2952	2953	2954	2955	2956
	2957	2958	2959	2960	2961	2962	2963	2964	2965	2966	2967	2968	2969	2970	2972
	2974	2975	2976	2977	2978	2979	2980	2981	2982	2983	2984	2985	2986	2987	2988
	2989	2990	2991	2992	2993	2994	2995	2996	2997	2998	2999	3000	3001	3002	3003
	3004	3005	3006	3007	3008	3009	3010	3011	3012	3013	3014	3015	3016	3017	3018
	3019	3020	3021	3022	3023	3024	3025	3026	3027	3028	3029	3030	3031	3032	3033
	3034	3035	3036	3037	3038	3039	3040	3041	3042	3043	3044	3045	3046	3047	3048
	3049	3050	3051	3052	3053	3054	3055	3056	3057	3058	3059	3060	3061	3062	3063
	3064	3065	3066	3067	3068	3069	3070	3071	3072	3073	3074	3076	3077	3078	3079
	3080	3081	3082	3083	3084	3085	3086	3087	3088	3089	3090	3091	3092	3093	3094
	3095	3096	3097	3098	3099	3100	3101	3102	3103	3104	3105	3106	3107	3108	3109
	3110	3111	3112	3113	3114	3115	3116	3117	3118	3119	3120	3121	3122	3123	3124
	3125	3126	3127	3128	3129	3130	3131	3132	3133	3134	3135	3136	3137	3138	3139
	3140	3141	3142	3143	3144	3145	3146	3147	3148	3149	3150	3151	3152	3154	3155
	3156	3157	3158	3159	3160	3161	3162	3163	3164	3165	3166	3167	3168	3169	3170
	3171	3172	3173	3174	3175	3176	3177	3178	3179	3180	3181	3182	3183	3184	3185
	3186	3187	3188	3189	3190	3191	3192	3193	3194	3196	3197	3199	3200	3201	3202
	3203	3204	3205	3206	3207	3208	3209	3210	3211	3212	3213	3214	3215	3216	3217
	3218	3219	3220	3221	3222	3223	3224	3226	3227	3228	3229	3230	3231	3232	3233

Cluster No.	Document Number														
	3234	3235	3236	3237	3238	3239	3240	3241	3242	3243	3244	3245	3246	3247	3248
	3249	3250	3251	3252	3253	3254	3255	3256	3257	3258	3259	3260	3261	3263	3265
	3266	3267	3268	3269	3270	3271	3272	3273	3274	3275	3276	3277	3278	3279	3280
	3281	3282	3283	3284	3285	3286	3287	3288	3289	3290	3291	3292	3293	3294	3295
	3296	3297	3298	3299	3300	3301	3302	3303	3304	3305	3306	3307	3308	3309	3310
	3311	3312	3313	3314	3315	3316	3317	3318	3319	3320	3321	3322	3323	3324	3325
	3326	3327	3328	3329	3330	3331	3332	3333	3334	3335	3336	3337	3338	3339	3340
	3341	3342	3343	3344	3345	3346	3347	3348	3349	3350	3351	3352	3353	3354	3355
	3356	3357	3358	3359	3360	3361	3362	3363	3364	3365	3366	3367	3368	3369	3370
	3371	3372	3373	3374	3375	3376	3377	3378	3379	3380	3381	3382	3383	3384	3385
	3386	3387	3388	3389	3390	3391	3392	3393	3394	3395	3396	3397	3398	3399	3400
	3401	3402	3403	3404	3405	3406	3407	3408	3409	3410	3411	3412	3413	3414	3415
	3416	3417	3418	3419	3420	3421	3422	3423	3424	3425	3426	3427	3428	3429	3430
	3431	3432	3433	3434	3435	3436	3437	3438	3439	3440	3441	3442	3443	3444	3445
	3446	3447	3451	3452	3453	3454	3455	3456	3458	3461	3463	3464	3465	3466	3467
	3468	3469	3470	3471	3472	3473	3474	3475	3476	3477	3478	3479	3480	3481	3482
	3483	3484	3485	3486	3487	3488	3489	3490	3491	3492	3493	3494	3495	3496	3497
	3498	3499	3500	3501	3502	3503	3504	3505	3506	3507	3508	3509	3510	3511	3512
	3513	3514	3515	3516	3517	3518	3519	3520	3521	3522	3523	3524	3525	3526	3527
	3528	3530	3531	3532	3533	3534	3535	3536	3537	3538	3539	3540	3541	3542	3543
	3544	3545	3546	3547	3548	3549	3550	3551	3553	3554	3555	3556	3558	3559	3560
	3561	3562	3563	3564	3565	3566	3567	3568	3569	3570	3571	3572	3573	3574	3575
	3576	3577	3578	3579	3580	3581	3582	3583	3584	3585	3586	3587	3588	3589	3590
	3591	3592	3593	3594	3595	3596	3597	3598	3599	3600	3601	3602	3603	3605	3606
	3607	3608	3609	3610	3611	3612	3613	3614	3615	3616	3617	3618	3619	3620	3621
	3622	3623	3624	3625	3626	3627	3628	3629	3630	3631	3632	3633	3634	3635	3636
	3637	3639	3640	3641	3642	3643	3644	3645	3646	3647	3648	3649	3650	3651	3652
	3653	3654	3655	3656	3657	3658	3659	3660	3661	3662	3663	3664	3665	3666	3667
	3669	3670	3671	3672	3673	3674	3675	3676	3677	3678	3679	3680	3681	3682	3683
	3684	3685	3686	3687	3688	3689	3690	3691	3692	3693	3694	3695	3696	3697	3698
	3699	3700	3701	3702	3703	3704	3705	3706	3707	3708	3709	3710	3711	3712	3713
	3714	3715	3716	3717	3718	3719	3720	3721	3722	3723	3724	3725	3726	3727	3728
	3729	3730	3731	3733	3734	3735	3736	3737	3738	3740	3741	3742	3743	3744	3745
	3746	3747	3748	3749	3750	3751	3752	3753	3754	3755	3756	3757	3758	3759	3760
	3761	3762	3763	3764	3765	3766	3767	3768	3769	3770	3771	3772	3773	3774	3775
	3776	3777	3778	3779	3780	3781	3782	3783	3784	3785	3786	3787	3788	3789	3790
	3791	3792	3793	3794	3795	3796	3797	3798	3799	3800	3801	3802	3803	3804	3805
	3806	3807	3809	3810	3811	3812	3813	3814	3815	3816	3817	3818	3819	3820	3821
	3822	3823	3824	3825	3826	3827	3828	3829	3830	3831	3832	3833	3834	3835	3836
	3837	3838	3839	3840	3841	3842	3843	3844	3845	3846	3847	3848	3849	3850	3851
	3852	3854	3855	3856	3857	3858	3859	3860	3861	3862	3863	3864	3865	3866	3867
	3869	3870	3871	3872	3873	3874	3875								
Cluster	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15

Cluster No.	Document Number														
#3	16	17	18	19	20	21	22	23	24	25	26	27	28	29	31
	32	33	34	36	37	38	39	40	41	42	43	44	45	46	47
	48	49	50	51	52	53	54	55	56	57	58	59	60	61	62
	63	64	65	66	67	68	69	70	71	72	73	74	75	76	77
	78	79	80	81	82	83	84	85	86	87	88	89	90	91	92
	93	95	96	97	98	100	102	103	104	105	106	107	109	110	111
	112	113	114	115	116	117	119	120	121	122	123	124	125	126	127
	128	129	130	131	132	133	134	135	136	137	138	139	140	141	142
	143	144	145	146	147	148	149	150	151	152	153	154	155	156	158
	159	160	161	162	163	164	165	166	167	169	170	171	172	173	174
	175	176	177	178	179	180	181	182	183	184	185	186	187	188	189
	190	191	192	193	194	195	196	197	198	199	200	201	202	203	204
	205	206	207	208	209	210	211	212	213	214	215	216	217	218	219
	220	221	222	223	224	225	226	228	230	232	234	235	236	237	238
	239	240	241	242	243	244	245	246	247	248	249	250	251	252	253
	254	255	257	258	259	260	261	262	263	264	265	266	267	268	269
	270	271	272	273	274	275	276	277	278	279	280	281	282	283	284
	285	286	287	288	289	290	291	292	293	294	295	296	297	298	299
	300	301	302	303	304	305	307	308	309	310	311	312	313	314	315
	316	317	318	319	320	321	322	323	324	325	326	327	328	329	330
	331	332	333	334	336	337	338	339	340	341	342	343	344	345	346
	348	349	350	351	352	353	354	355	356	357	358	359	360	361	362
	363	364	365	366	368	369	370	371	372	373	374	375	376	377	378
	379	380	381	382	383	384	385	386	387	388	390	391	392	393	394
	395	396	397	398	399	400	401	402	403	404	405	406	407	408	409
	410	411	412	413	417	418	419	420	421	422	424	425	426	427	429
	430	431	432	433	434	435	436	437	438	440	441	442	443	444	445
	446	447	448	449	450	451	452	453	454	455	456	457	458	459	460
	461	463	464	465	466	467	468	469	470	472	473	474	475	476	477
	478	479	480	481	482	483	484	485	486	487	488	489	490	491	493
	494	495	496	497	498	499	500	501	502	503	504	505	506	507	508
	509	510	511	512	513	514	515	516	517	518	519	520	521	522	523
	524	525	526	527	528	529	530	531	532	533	535	536	537	538	539
	540	541	542	543	544	545	546	548	549	550	551	552	553	554	555
	557	558	559	560	561	562	563	564	565	566	567	568	569	570	571
	572	573	574	575	576	577	578	579	580	581	582	583	584	585	586
	587	588	589	590	591	592	593	594	595	596	597	598	599	600	601
	602	604	605	606	607	608	609	610	611	612	613	614	615	616	617
	618	619	620	621	622	623	624	625	627	628	629	630	631	632	633
	634	635	636	637	638	639	640	641	642	643	644	645	646	647	648
	649	650	651	652	653	654	655	656	657	658	659	660	661	662	664
	666	667	668	669	670	671	672	673	674	675	676	677	678	679	680
	681	682	683	684	685	686	687	688	689	690	691	692	693	694	695

Cluster No.	Document Number														
	696	697	698	699	700	701	702	703	704	705	706	707	708	709	710
	711	712	713	714	715	717	718	719	720	721	722	723	724	725	726
	727	728	729	730	731	732	733	734	735	736	737	738	739	740	741
	742	744	745	746	747	749	750	751	752	753	754	755	756	757	758
	759	760	761	762	763	764	765	766	767	768	769	770	771	772	773
	774	775	776	777	778	779	780	781	782	783	784	785	786	787	788
	790	791	792	793	794	795	796	797	799	800	801	802	803	804	805
	806	807	808	809	810	811	812	813	814	815	816	817	818	819	820
	821	822	823	824	825	826	827	828	829	830	831	832	833	834	835
	836	837	838	839	840	841	842	843	844	845	846	847	848	849	850
	851	852	853	854	855	856	857	858	859	861	862	863	864	865	866
	867	868	869	870	871	872	873	874	875	876	877	878	879	880	881
	882	883	884	885	886	887	889	890	892	893	894	895	896	897	898
	899	901	902	903	904	905	906	907	908	910	911	912	913	914	915
	916	917	918	919	920	921	922	923	924	925	926	927	928	929	930
	931	932	933	934	935	936	937	938	939	940	941	942	943	944	945
	946	947	948	949	950	951	952	953	954	955	956	957	958	959	960
	961	962	963	964	965	966	967	968	969	970	971	972	973	974	975
	977	978	979	980	981	982	983	984	985	986	987	988	989	990	992
	993	994	995	996	997	998	999	1000	1001	1002	1003	1004	1005	1006	1007
	1008	1009	1011	1012	1013	1014	1015	1016	1017	1018	1019	1020	1021	1022	1023
	1024	1026	1028	1029	1030	1031	1033	1034	1035	1036	1038	1039	1040	1043	1044
	1045	1047	1048	1049	1050	1051	1052	1053	1054	1055	1056	1057	1058	1060	1062
	1063	1064	1066	1067	1068	1069	1070	1071	1072	1073	1074	1075	1076	1077	1079
	1080	1081	1083	1085	1086	1087	1088	1090	1091	1092	1093	1094	1095	1096	1097
	1098	1099	1100	1101	1102	1103	1104	1105	1106	1107	1108	1109	1110	1111	1112
	1114	1116	1117	1118	1119	1120	1121	1122	1123	1124	1125	1126	1127	1128	1129
	1130	1131	1132	1133	1134	1135	1136	1138	1139	1140	1141	1142	1143	1144	1145
	1147	1148	1149	1151	1152	1154	1156	1157	1158	1159	1160	1161	1162	1163	1165
	1166	1167	1168	1169	1170	1171	1172	1173	1174	1175	1176	1177	1178	1179	1180
	1181	1183	1185	1186	1187	1188	1189	1190	1191	1192	1193	1194	1196	1197	1198
	1199	1200	1201	1202	1203	1204	1205	1206	1207	1208	1209	1210	1211	1212	1213
	1215	1216	1217	1218	1219	1220	1221	1222	1223	1224	1225	1226	1227	1229	1230
	1231	1232	1233	1234	1235	1236	1237	1238	1240	1241	1242	1243	1244	1245	1246
	1247	1248	1249	1250	1252	1253	1254	1255	1256	1257	1258	1259	1260	1261	1262
	1263	1265	1266	1267	1269	1270	1271	1272	1273	1274	1275	1276	1278	1279	1280
	1282	1283	1284	1285	1286	1287	1288	1289	1291	1293	1294	1295	1296	1298	1299
	1301	1303	1304	1305	1306	1307	1308	1309	1311	1312	1313	1314	1315	1317	1318
	1320	1322	1324	1325	1326	1328	1330	1331	1332	1333	1334	1335	1336	1337	1338
	1339	1340	1341	1342	1343	1344	1345	1346	1347	1348	1349	1350	1351	1352	1353
	1354	1355	1356	1357	1358	1359	1360	1361	1362	1363	1364	1365	1366	1368	1369
	1370	1371	1373	1374	1376	1377	1378	1379	1380	1381	1382	1383	1386	1387	1388
	1389	1390	1391	1393	1394	1395	1396	1398	1399	1400	1401	1402	1403	1404	1405

Cluster No.	Document Number														
	1406	1407	1408	1409	1410	1411	1412	1413	1414	1416	1417	1418	1419	1420	1421
	1422	1424	1425	1426	1427	1428	1429	1430	1431	1432	1433	1434	1435	1436	1438
	1439	1440	1441	1443	1444	1445	1446	1492	1576	1630	1813	2043	2163	2190	2320
	2413	2546	2832	3153	3195	3448	3449	3457	3638	3732	3739	3868			

Data set ‘YahooK1’ (= misclassified document)

Cluster No.	Document Number														
Cluster # 1	495	496	497	498	499	500	502	504	505	507	508	510	511	517	518
	519	525	526	534	539	544	545	546	547	548	555	556	558	559	561
	562	564	565	567	576	577	578	580	581	582	583	584	585	586	588
	591	606	609	610	611	612	613	614	615	616	617	618	619	620	621
	626	627	629	630	631	632	633	634	635	636	644	647	648	649	650
	651	653	664	665	666	669	677	678	681	683	689	690	699	703	710
	711	712	713	714	715	716	719	720	721	722	723	724	737	738	741
	742	743	745	752	753	757	758	762	763	765	766	767	770	771	772
	776	789	790	794	795	796	797	798	799	800	801	803	804	807	808
	811	812	813	814	815	816	824	825	827	828	829	830	831	832	834
	835	836	837	838	839	840	841	844	848	851	857	862	866	868	869
	870	871	877	878	879	880	881	882	883	884	885	886	887	889	893
	894	898	913	934	935	936	937	942	943	944	945	946	947	948	949
	952	956	957	958	959	960	961	962	963	965	966	967	968	973	976
	977	978	979	986	987	995	996	1002	1003	1011	1012	1015	1016	1017	1018
	1021	1026	1027	1028	1029	1030	1039	1040	1041	1042	1048	1049	1057	1058	1059
	1060	1062	1063	1064	1065	1067	1068	1069	1070	1071	1072	1073	1074	1075	1076
	1077	1079	1080	1084	1086	1087	1089	1090	1091	1094	1099	1100	1101	1102	1104
	1105	1106	1107	1108	1109	1110	1112	1121	1122	1123	1124	1132	1136	1139	1142
	1143	1148	1149	1150	1151	1152	1153	1171	1174	1175	1176	1177	1178	1179	1180
	1181	1182	1183	1184	1186	1187	1189	1193	1198	1201	1209	1211	1212	1213	1214
	1215	1219	1220	1228	1229	1230	1231	1233	1234	1235	1236	1237	1238	1268	1269
	1270	1271	1272	1273	1274	1275	1278	1279	1280	1281	1282	1283	1284	1293	1294
	1298	1302	1306	1311	1312	1313	1314	1315	1316	1317	1318	1319	1320	1328	1331
	1333	1335	1336	1337	1338	1342	1345	1349	1350	1351	1352	1353	1354	1355	1362
	1364	1366	1367	1368	1369	1372	1376	1394	1397	1398	1399	1400	1401	1402	1403
	1404	1405	1409	1435	1440	1441	1442	1445	1449	1460	1461	1462	1468	1470	1471
	1472	1473	1474	1475	1476	1477	1478	1479	1486	1487	1488	1490	1491	1492	1495
	1497	1512	1513	1514	1517	1518	1521	1522	1523	1524	1525	1527	1528	1530	1531
	1536	1537	1538	1543	1544	1546	1547	1548	1549	1553	1554	1556	1557	1559	1570
	1571	1572	1575	1576	1577	1578	1579	1580	1581	1583	1584	1585	1586	1593	1594
	1595	1596	1611	1614	1615	1616	1617	1621	1623	1624	1631	1632	1633	1634	1635
	1636	1637	1638	1639	1642	1643	1644	1645	1646	1647	1648	1649	1650	1651	1654
	1655	1657	1658	1659	1660	1661	1662	1663	1664	1669	1676	1681	1687	1689	1690

Cluster No.	Document Number														
	1691	1692	1694	1695	1696	1697	1698	1699	1700	1701	1704	1705	1710	1713	1737
	1738	1739	1742	1743	1744	1745	1746	1747	1748	1749	1750	1751	1752	1753	1754
	1755	1756	1766	1767	1770	1781	1783	1784	1785	1786	1793	1799	1805	1806	1807
	1820	1821	1822	1823	1825	1838	1840	1841	1842	1843	1849	1850	1851	1852	1855
	1857	1859	1860	1866	1867	1868	1869	1870	1875	1880	1883	2203	2208		
Cluster # 2	1	2	3	4	5	6	7	8	9	10	12	13	14	15	16
	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32
	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48
	49	50	51	52	53	54	55	56	58	59	60	61	62	63	64
	65	66	67	68	69	70	71	72	73	74	75	77	78	79	80
	81	82	83	84	85	86	87	88	89	90	91	92	93	94	95
	96	97	98	101	102	103	104	105	106	107	108	109	110	111	112
	113	114	115	118	119	120	121	122	124	126	127	128	129	130	131
	132	133	134	135	136	137	138	139	140	141	142	143	144	145	146
	147	148	150	151	152	153	155	156	157	158	159	160	161	162	163
	164	165	166	168	169	170	171	172	173	174	175	176	177	178	179
	180	181	184	185	186	187	188	189	190	191	192	193	194	195	196
	197	198	199	200	201	202	203	204	205	206	207	208	209	210	211
	212	213	214	215	216	217	218	219	220	221	222	223	224	225	226
	229	230	231	232	233	234	235	236	237	238	239	240	241	242	243
	244	245	246	247	248	249	250	251	252	253	254	255	257	258	259
	260	261	262	263	264	265	266	267	268	269	270	272	273	274	275
	276	277	278	279	280	281	282	283	284	285	286	287	288	289	290
	291	292	293	294	295	296	297	298	299	301	302	303	304	305	306
	307	308	309	311	312	313	314	315	316	317	318	319	320	321	322
	323	324	325	326	327	328	329	330	331	332	333	334	336	337	338
	339	340	341	342	343	344	346	347	348	349	350	351	352	353	354
	355	356	357	358	359	360	361	362	363	364	365	366	367	368	369
	370	371	372	373	374	375	376	377	378	379	380	381	382	383	384
	385	386	387	388	389	390	391	392	393	394	395	396	397	398	399
	401	402	403	405	406	407	408	409	410	411	412	413	414	415	416
	417	418	419	420	421	422	423	424	425	426	427	428	429	430	431
	432	433	434	435	436	437	438	439	441	442	443	444	445	446	447
	448	449	450	451	452	453	454	455	457	458	459	460	461	462	463
	464	466	467	468	469	470	471	472	473	475	476	477	478	479	480
	481	482	483	484	485	486	487	488	489	490	491	492	493	494	506
	744	852	1303	1332	1656	1732	2309	2335							
Cluster #3	76	99	100	116	117	123	125	149	154	167	182	183	228	256	271
	310	335	345	400	404	465	474	528	536	543	560	563	602	608	623
	624	637	638	668	675	676	692	693	694	701	717	718	725	736	739
	740	746	748	749	751	755	759	781	785	791	802	818	842	843	845
	846	849	850	856	858	859	860	861	888	895	896	900	902	903	904
	914	938	950	953	969	970	991	998	1019	1036	1037	1044	1046	1144	1146

Cluster No.	Document Number														
	1147	1163	1164	1165	1190	1191	1192	1207	1208	1216	1217	1240	1241	1242	1256
	1257	1258	1285	1288	1289	1290	1291	1295	1296	1300	1305	1309	1343	1344	1359
	1370	1371	1373	1375	1410	1411	1412	1413	1414	1415	1416	1417	1418	1419	1421
	1423	1424	1425	1426	1427	1428	1433	1434	1436	1437	1438	1446	1447	1448	1469
	1482	1483	1484	1493	1494	1504	1509	1515	1529	1539	1540	1551	1587	1588	1589
	1590	1597	1600	1601	1608	1609	1613	1622	1627	1628	1630	1652	1653	1665	1671
	1672	1673	1674	1677	1678	1679	1680	1682	1693	1702	1703	1706	1707	1714	1715
	1716	1717	1718	1719	1720	1721	1722	1723	1724	1725	1726	1733	1734	1735	1740
	1741	1768	1769	1771	1772	1773	1774	1775	1776	1777	1778	1779	1780	1787	1788
	1790	1795	1796	1797	1800	1801	1802	1803	1804	1808	1809	1810	1811	1824	1826
	1829	1830	1831	1832	1833	1834	1835	1836	1848	1853	1854	1856	1871	1872	1873
	1876	1878	1879	1881	1882	2207									
Cluster #4	11	17	33	57	503	512	529	533	537	538	540	541	542	553	557
	575	579	589	640	643	672	686	688	695	702	730	768	786	853	874
	897	912	915	924	925	926	927	928	929	930	931	932	954	964	1043
	1066	1083	1166	1169	1196	1197	1204	1248	1250	1255	1276	1277	1287	1301	1365
	1374	1444	1481	1498	1500	1503	1550	1558	1736	1798	1874	1884	1885	1886	1887
	1888	1889	1890	1891	1892	1893	1894	1895	1896	1897	1898	1899	1900	1901	1902
	1903	1904	1905	1906	1907	1908	1909	1910	1911	1912	1913	1914	1915	1916	1917
	1918	1919	1920	1921	1922	1923	1924	1925	1926	1927	1928	1929	1930	1931	1932
	1933	1934	1935	1936	1937	1938	1939	1940	1941	1942	1943	1944	1945	1946	1947
	1948	1949	1950	1951	1952	1953	1954	1955	1956	1957	1958	1959	1960	1961	1962
	1963	1964	1965	1966	1967	1968	1969	1970	1971	1972	1973	1974	1975	1976	1977
	1978	1979	1980	1981	1982	1983	1984	1985	1986	1987	1988	1989	1990	1991	1992
	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007
	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022
	2023	2024	2025	2026	2027	2028	2030	2031	2032	2033	2034	2035	2036	2037	2038
	2039	2040	2041	2042	2043	2044	2045	2046	2047	2048	2049	2050	2051	2052	2053
	2054	2055	2056	2057	2058	2059	2060	2061	2062	2063	2064	2065	2066	2067	2068
	2069	2070	2071	2073	2074	2075	2076	2077	2078	2079	2080	2081	2082	2083	2084
	2085	2086	2087	2088	2089	2090	2091	2092	2093	2094	2095	2096	2097	2098	2099
	2100	2101	2102	2103	2104	2105	2106	2107	2108	2109	2110	2111	2112	2113	2114
	2115	2116	2117	2118	2119	2120	2121	2122	2123	2124	2125	2126	2127	2128	2129
	2130	2131	2132	2133	2134	2135	2136	2137	2138	2217	2235	2236	2237	2238	2261
	2308														
Cluster #5	227	300	440	456	522	530	594	595	596	598	622	642	670	682	687
	700	727	728	729	731	732	733	734	735	747	754	778	784	787	788
	833	847	855	890	892	901	906	974	975	992	999	1000	1001	1038	1052
	1098	1130	1141	1145	1188	1199	1200	1202	1227	1243	1244	1246	1247	1249	1253
	1292	1297	1304	1307	1308	1341	1443	1499	1501	1507	1582	1591	1592	1598	1602
	1603	1607	1612	1668	1675	1728	1729	1730	1791	1792	1877	2029	2072	2139	2140
	2141	2142	2143	2144	2145	2146	2147	2148	2149	2150	2151	2152	2153	2154	2155
	2156	2157	2158	2159	2160	2161	2162	2163	2164	2165	2166	2167	2168	2169	2170

Cluster No.	Document Number														
	2171	2172	2173	2174	2175	2176	2177	2178	2179	2180	2181	2182	2183	2184	2185
	2186	2187	2188	2189	2190	2191	2192	2193	2194	2195	2196	2197	2198	2199	2200
	2201	2202	2204	2205	2206	2209	2210	2211	2212	2213	2214	2215	2216	2218	2219
	2220	2221	2222	2223	2224	2225	2226	2227	2228	2229	2230	2231	2232	2233	2234
	2239	2240	2241	2242	2243	2244	2245	2246	2247	2248	2249	2250	2251	2252	2253
	2254	2255	2256	2257	2258	2259	2260	2262	2263	2264	2265	2266	2267	2268	2269
	2270	2271	2272	2273	2274	2275	2276	2277	2278	2279	2280	2281	2282	2283	2284
	2285	2286	2287	2288	2289	2290	2291	2292	2293	2294	2295	2296	2297	2298	2299
	2300	2301	2302	2303	2304	2305	2306	2307	2310	2311	2312	2313	2314	2315	2316
	2317	2318	2319	2320	2321	2322	2323	2324	2325	2326	2327	2328	2329	2330	2331
	2332	2333	2334	2336	2337	2338	2339	2340							
Cluster # 6	501	509	513	514	515	516	520	521	523	524	527	531	532	535	549
	550	551	552	554	566	568	569	570	571	572	573	574	587	590	592
	593	597	599	600	601	603	604	605	607	625	628	639	641	645	646
	652	654	655	656	657	658	659	660	661	662	663	667	671	673	674
	679	680	684	685	691	696	697	698	704	705	706	707	708	709	726
	750	756	760	761	764	769	773	774	775	777	779	780	782	783	792
	793	805	806	809	810	817	819	820	821	822	823	826	854	863	864
	865	867	872	873	875	876	891	899	905	907	908	909	910	911	916
	917	918	919	920	921	922	923	933	939	940	941	951	955	971	972
	980	981	982	983	984	985	988	989	990	993	994	997	1004	1005	1006
	1007	1008	1009	1010	1013	1014	1020	1022	1023	1024	1025	1031	1032	1033	1034
	1035	1045	1047	1050	1051	1053	1054	1055	1056	1061	1078	1081	1082	1085	1088
	1092	1093	1095	1096	1097	1103	1111	1113	1114	1115	1116	1117	1118	1119	1120
	1125	1126	1127	1128	1129	1131	1133	1134	1135	1137	1138	1140	1154	1155	1156
	1157	1158	1159	1160	1161	1162	1167	1168	1170	1172	1173	1185	1194	1195	1203
	1205	1206	1210	1218	1221	1222	1223	1224	1225	1226	1232	1239	1245	1251	1252
	1254	1259	1260	1261	1262	1263	1264	1265	1266	1267	1286	1299	1310	1321	1322
	1323	1324	1325	1326	1327	1329	1330	1334	1339	1340	1346	1347	1348	1356	1357
	1358	1360	1361	1363	1377	1378	1379	1380	1381	1382	1383	1384	1385	1386	1387
	1388	1389	1390	1391	1392	1393	1395	1396	1406	1407	1408	1420	1422	1429	1430
	1431	1432	1439	1450	1451	1452	1453	1454	1455	1456	1457	1458	1459	1463	1464
	1465	1466	1467	1480	1485	1489	1496	1502	1505	1506	1508	1510	1511	1516	1519
	1520	1526	1532	1533	1534	1535	1541	1542	1545	1552	1555	1560	1561	1562	1563
	1564	1565	1566	1567	1568	1569	1573	1574	1599	1604	1605	1606	1610	1618	1619
	1620	1625	1626	1629	1640	1641	1666	1667	1670	1683	1684	1685	1686	1688	1708
	1709	1711	1712	1727	1731	1757	1758	1759	1760	1761	1762	1763	1764	1765	1782
	1789	1794	1812	1813	1814	1815	1816	1817	1818	1819	1827	1828	1837	1839	1844
	1845	1846	1847	1858	1861	1862	1863	1864	1865						

E.2. PCA

Data set ‘Binary2’ (= misclassified document)

Cluster No.	Document Number														
cluster #1	1	2	3	4	5	6	7	8	9	10	14	15	16	17	18
	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33
	34	35	36	37	38	39	41	42	43	44	46	47	48	49	51
	52	53	54	55	56	57	58	59	60	61	62	63	64	65	66
	67	68	69	70	71	72	73	74	75	76	77	78	79	80	81
	82	83	84	85	86	87	88	89	90	92	93	94	95	96	97
	99	101	102	103	104	105	106	107	108	109	110	111	112	113	114
	115	116	117	118	119	120	121	122	123	124	125	126	127	128	129
	130	131	132	133	134	135	136	137	138	139	140	141	142	143	144
	145	146	147	148	149	150	151	152	153	154	155	156	157	158	159
	160	161	162	163	164	165	166	167	168	169	170	171	172	173	174
	175	176	177	178	179	180	181	182	183	184	185	186	187	188	189
	190	191	192	193	194	195	196	197	198	199	200	201	202	203	204
	205	206	207	208	209	210	211	212	213	214	215	216	217	218	219
	220	221	222	223	224	225	226	227	228	229	230	231	232	233	234
	235	236	237	238	239	240	241	242	243	244	245	246	247	248	249
	250	258	260	278	289	297	298	313	314	322	341	378	380	388	389
	397	399	402	403	404	415	419	434	440	442	465	474	482	488	498
cluster #2	11	12	13	40	45	50	91	98	100	251	252	253	254	255	256
	257	259	261	262	263	264	265	266	267	268	269	270	271	272	273
	274	275	276	277	279	280	281	282	283	284	285	286	287	288	290
	291	292	293	294	295	296	299	300	301	302	303	304	305	306	307
	308	309	310	311	312	315	316	317	318	319	320	321	323	324	325
	326	327	328	329	330	331	332	333	334	335	336	337	338	339	340
	342	343	344	345	346	347	348	349	350	351	352	353	354	355	356
	357	358	359	360	361	362	363	364	365	366	367	368	369	370	371
	372	373	374	375	376	377	379	381	382	383	384	385	386	387	390
	391	392	393	394	395	396	398	400	401	405	406	407	408	409	410
	411	412	413	414	416	417	418	420	421	422	423	424	425	426	427
	428	429	430	431	432	433	435	436	437	438	439	441	443	444	445
	446	447	448	449	450	451	452	453	454	455	456	457	458	459	460
	461	462	463	464	466	467	468	469	470	471	472	473	475	476	477
	478	479	480	481	483	484	485	486	487	489	490	491	492	493	494
	495	496	497	499	500										

Data set ‘Multi5’ (▭ = misclassified document)

Cluster No.	Document Number														
cluster #1	46	291	301	302	303	304	305	306	307	308	309	310	311	312	313
	314	315	316	317	318	319	320	321	322	323	324	325	326	327	328
	329	330	331	332	333	334	335	336	337	338	339	340	341	342	343
	344	345	346	347	348	349	350	351	352	353	354	355	357	358	359
	360	361	362	363	364	365	366	368	369	370	371	372	373	374	375
	376	377	378	379	380	381	382	383	384	385	386	387	388	389	390
	391	392	393	394	395	396	397	398	399	400					
cluster #2	401	402	403	404	405	406	407	408	409	410	411	412	413	414	415
	416	417	418	419	420	421	422	423	424	425	426	427	428	429	430
	431	432	433	434	435	436	437	438	439	440	441	442	443	444	445
	446	447	448	449	450	451	452	453	454	455	456	457	458	460	461
	462	463	464	465	466	467	468	469	470	471	472	473	474	475	476
	477	478	479	480	481	482	483	484	485	486	487	488	489	490	491
	492	493	494	495	496	497	498	499	500						
cluster #3	85	201	202	203	204	205	206	207	208	209	210	211	212	213	214
	215	216	217	218	219	220	221	222	223	224	225	226	227	228	229
	230	231	232	233	234	235	236	237	238	239	240	241	242	243	244
	245	246	247	248	249	251	252	253	254	255	256	257	258	259	260
	261	262	263	264	265	266	267	268	269	270	271	272	273	274	275
	277	278	279	281	282	283	284	285	286	287	288	289	290	292	294
	295	296	298	299	300										
cluster #4	101	102	103	104	105	106	107	108	109	110	111	112	113	114	115
	116	117	118	119	120	121	122	123	124	125	126	127	128	129	130
	131	132	133	134	135	136	137	138	139	140	141	142	143	144	145
	146	147	148	149	150	151	152	153	154	155	156	157	158	159	160
	161	162	163	164	165	166	167	168	169	170	172	173	175	176	177
	178	179	180	181	182	183	184	186	187	188	189	190	191	192	193
	194	195	196	197	198	199	200	276	293						
cluster #5	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45
	47	48	49	50	51	52	53	54	55	56	57	58	59	60	61
	62	63	64	65	66	67	68	69	70	71	72	73	74	75	76
	77	78	79	80	81	82	83	84	86	87	88	89	90	91	92
	93	94	95	96	97	98	99	100	171	174	185	250	280	297	356
	367	459													

Data set ‘Multi10’ (= misclassified document)

Cluster No.	Document Number														
cluster #1	232	233	248	251	252	253	254	255	256	257	258	259	260	261	262
	263	264	265	268	269	270	271	272	273	274	275	276	277	278	279
	280	281	282	283	284	285	286	287	288	289	290	291	292	293	294
	295	296	297	298	299	300	301	308	343	352					
cluster #2	29	52	57	83	101	102	103	104	105	106	108	109	111	112	113
	115	116	117	118	119	120	121	122	123	124	125	126	127	128	129
	131	132	133	134	135	136	137	138	139	140	141	142	143	144	146
	147	148	149	150	205	208	227	235	307	353	375	420	444		
cluster #3	145	201	202	203	204	206	209	210	211	212	213	214	215	216	217
	218	219	220	221	222	223	224	225	226	228	229	230	231	234	236
	237	238	240	241	243	244	245	246	247	249	250	355			
cluster #4	23	24	239	303	310	311	313	315	318	322	325	326	329	330	331
	332	333	334	335	336	337	339	340	341	342	344	349	357	367	369
	371	372	400	421	425	431	439	440	442	443	445	450			
cluster #5	18	32	93	151	152	153	154	155	156	157	158	159	160	161	162
	163	164	165	166	167	168	169	170	171	172	173	174	175	176	177
	178	179	180	181	182	183	185	186	187	188	189	190	191	192	193
	194	195	196	197	198	199	200	207	242	302	304	305	306	309	312
	314	358	377												
cluster #6	5	114	130	316	317	319	327	338	347	348	351	354	356	359	360
	363	364	365	366	368	370	373	374	376	378	379	380	381	382	383
	384	385	386	387	388	389	390	391	392	393	394	395	396	397	398
	399	435													
cluster #7	39	88	107	110	321	423	446	451	452	453	454	455	456	457	458
	459	460	461	462	463	464	465	466	467	468	469	470	471	472	473
	474	475	476	477	478	479	480	481	482	483	484	485	486	487	488
	489	490	491	492	493	494	495	496	497	498	499	500			
cluster #8	51	53	54	55	56	58	59	60	61	62	63	64	65	66	67
	68	69	70	71	72	73	74	75	76	77	78	79	80	81	82
	84	85	86	87	89	90	91	92	94	95	96	97	98	99	100
	184	266	323	324	328	345	346	350	361	362					
cluster #9	35	267	320	401	402	403	404	405	406	407	408	409	410	411	412
	413	414	415	416	417	418	419	422	424	426	427	428	429	430	432
	433	434	436	437	438	441	447	448	449						
cluster #10	1	2	3	4	6	7	8	9	10	11	12	13	14	15	16
	17	19	20	21	22	25	26	27	28	30	31	33	34	36	37
	38	40	41	42	43	44	45	46	47	48	49	50			

Data set ‘Classic3’ (▭ = misclassified document)

Cluster No.	Document Number														
cluster #1	229	233	306	335	347	367	439	462	492	519	547	663	716	789	991
	1010	1027	1037	1082	1084	1150	1214	1228	1239	1249	1277	1319	1323	1372	1384
	1385	1397	1415	1447	1448	1449	1450	1451	1452	1453	1454	1455	1456	1457	1458
	1459	1460	1461	1462	1463	1464	1465	1466	1467	1468	1469	1470	1471	1472	1473
	1474	1475	1476	1477	1478	1479	1480	1481	1482	1483	1484	1485	1486	1487	1488
	1489	1490	1491	1493	1494	1495	1496	1497	1498	1499	1500	1501	1502	1503	1504
	1505	1506	1507	1508	1509	1510	1511	1512	1513	1514	1515	1516	1517	1518	1519
	1520	1521	1522	1523	1524	1525	1526	1527	1528	1529	1530	1531	1532	1533	1534
	1535	1536	1537	1538	1539	1540	1541	1542	1543	1544	1545	1546	1547	1548	1549
	1550	1551	1552	1553	1554	1555	1556	1557	1558	1559	1560	1561	1562	1563	1564
	1565	1566	1567	1568	1569	1570	1571	1572	1573	1574	1575	1577	1578	1579	1580
	1581	1582	1583	1584	1585	1586	1587	1588	1590	1591	1592	1593	1594	1595	1596
	1597	1598	1599	1600	1601	1602	1603	1604	1605	1606	1607	1608	1609	1610	1611
	1612	1613	1614	1615	1616	1617	1618	1619	1620	1621	1622	1623	1624	1625	1626
	1627	1628	1629	1630	1631	1632	1633	1634	1635	1636	1637	1638	1639	1640	1641
	1642	1643	1644	1645	1646	1647	1648	1649	1650	1651	1652	1653	1654	1655	1656
	1657	1658	1659	1660	1661	1662	1663	1664	1665	1666	1667	1668	1669	1670	1671
	1672	1673	1674	1675	1676	1677	1678	1679	1680	1681	1682	1683	1684	1685	1686
	1687	1688	1689	1690	1691	1692	1693	1694	1695	1696	1697	1698	1699	1700	1701
	1702	1703	1704	1705	1706	1707	1708	1709	1710	1711	1712	1713	1714	1715	1716
	1717	1718	1719	1720	1721	1722	1723	1724	1725	1726	1727	1728	1729	1730	1731
	1732	1733	1734	1735	1736	1737	1738	1739	1740	1741	1742	1743	1744	1745	1746
	1747	1748	1749	1750	1751	1752	1753	1754	1755	1756	1757	1758	1759	1760	1761
	1762	1763	1764	1765	1766	1767	1768	1769	1770	1771	1772	1773	1774	1775	1776
	1777	1778	1779	1780	1781	1782	1783	1784	1785	1786	1787	1788	1789	1790	1791
	1792	1793	1794	1795	1796	1797	1798	1799	1800	1801	1802	1803	1804	1805	1806
	1807	1808	1809	1810	1811	1812	1814	1815	1816	1817	1818	1819	1820	1821	1822
	1823	1824	1825	1826	1827	1828	1829	1830	1831	1832	1833	1834	1835	1836	1837
	1838	1839	1840	1841	1842	1843	1844	1845	1846	1847	1848	1849	1850	1851	1852
	1853	1854	1855	1856	1857	1858	1859	1860	1861	1862	1863	1864	1865	1866	1867
	1868	1869	1870	1871	1872	1873	1874	1875	1876	1877	1878	1879	1880	1881	1882
	1883	1884	1885	1886	1887	1888	1889	1890	1891	1892	1893	1894	1895	1896	1897
	1898	1899	1900	1901	1902	1903	1904	1905	1906	1907	1908	1909	1910	1911	1912
	1913	1914	1915	1916	1917	1918	1919	1920	1921	1922	1923	1924	1925	1926	1927
	1928	1929	1930	1931	1932	1934	1935	1936	1937	1938	1939	1940	1941	1942	1943
	1944	1945	1946	1947	1948	1949	1950	1951	1952	1953	1954	1955	1956	1957	1958
	1959	1960	1961	1962	1963	1964	1965	1966	1967	1968	1969	1970	1971	1972	1973
	1974	1975	1976	1977	1978	1979	1980	1981	1982	1983	1984	1985	1986	1987	1988
	1989	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003
	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018
	2019	2020	2021	2022	2023	2024	2025	2026	2027	2028	2029	2030	2031	2032	2033

Cluster No.	Document Number														
	2034	2035	2036	2037	2038	2039	2040	2041	2042	2043	2044	2045	2046	2047	2048
	2049	2050	2051	2052	2053	2054	2055	2056	2057	2058	2059	2060	2062	2063	2064
	2065	2066	2067	2068	2069	2070	2071	2072	2073	2074	2075	2076	2077	2078	2079
	2080	2081	2082	2083	2084	2085	2086	2087	2088	2089	2090	2091	2092	2093	2095
	2096	2097	2098	2099	2100	2101	2102	2103	2104	2105	2106	2107	2108	2109	2110
	2111	2112	2113	2114	2115	2116	2117	2118	2119	2120	2121	2122	2123	2124	2125
	2126	2127	2128	2129	2130	2131	2132	2133	2134	2135	2136	2137	2138	2139	2140
	2141	2142	2143	2144	2145	2146	2147	2148	2149	2150	2151	2152	2153	2154	2155
	2156	2157	2158	2159	2160	2162	2164	2165	2166	2167	2168	2169	2170	2171	2172
	2173	2174	2175	2176	2177	2178	2179	2180	2181	2182	2183	2184	2185	2186	2187
	2188	2189	2191	2192	2193	2194	2195	2196	2197	2198	2199	2200	2201	2202	2203
	2204	2205	2206	2207	2208	2209	2210	2211	2212	2213	2214	2215	2216	2217	2218
	2219	2220	2221	2222	2223	2224	2225	2226	2227	2228	2229	2230	2231	2232	2233
	2234	2235	2236	2237	2238	2239	2240	2241	2242	2243	2244	2245	2246	2247	2248
	2249	2250	2251	2252	2253	2254	2255	2256	2257	2258	2259	2260	2261	2262	2263
	2264	2265	2266	2267	2268	2269	2270	2271	2272	2273	2274	2275	2276	2277	2278
	2279	2280	2281	2282	2283	2284	2285	2286	2287	2288	2289	2290	2291	2292	2293
	2294	2295	2296	2297	2298	2299	2300	2301	2302	2303	2304	2305	2306	2307	2308
	2309	2310	2311	2312	2313	2314	2315	2316	2317	2318	2319	2320	2321	2322	2323
	2324	2325	2326	2327	2328	2329	2330	2331	2332	2333	2334	2335	2336	2337	2338
	2339	2340	2341	2342	2343	2344	2345	2346	2347	2348	2349	2350	2351	2352	2353
	2354	2355	2356	2357	2358	2359	2360	2361	2362	2363	2364	2365	2366	2367	2368
	2369	2370	2371	2372	2373	2374	2375	2376	2377	2378	2379	2380	2381	2382	2383
	2384	2385	2386	2387	2388	2389	2390	2391	2392	2393	2394	2395	2396	2397	2398
	2399	2400	2401	2402	2404	2405	2406	2407	2408	2409	2410	2411	2412	2414	2415
	2416	2417	2418	2419	2420	2421	2422	2423	2424	2425	2426	2427	2428	2429	2430
	2431	2432	2433	2434	2435	2436	2437	2438	2439	2440	2441	2442	2443	2444	2445
	2446	2447	2448	2449	2450	2451	2452	2453	2454	2455	2456	2457	2458	2459	2460
	2461	2462	2463	2464	2465	2466	2467	2468	2469	2470	2471	2472	2473	2474	2475
	2491	2492	2493	2494	2495	2496	2497	2498	2499	2500	2501	2502	2503	2504	2505
	2506	2507	2508	2509	2510	2511	2512	2513	2514	2515	2516	2517	2518	2519	2520
	2521	2522	2523	2524	2525	2526	2527	2528	2529	2530	2531	2532	2533	2534	2535
	2536	2537	2538	2539	2540	2541	2542	2543	2544	2545	2547	2548	2549	2550	2551
	2552	2553	2554	2555	2556	2557	2558	2559	2560	2561	2562	2563	2564	2565	2566
	2567	2568	2569	2570	2571	2572	2573	2574	2575	2576	2577	2578	2579	2580	2581
	2582	2583	2584	2585	2586	2587	2588	2589	2590	2591	2592	2593	2594	2595	2596
	2597	2598	2599	2600	2601	2602	2603	2604	2605	2606	2607	2608	2609	2610	2611
	2612	2613	2614	2615	2616	2617	2618	2619	2620	2621	2622	2623	2624	2625	2626
	2627	2628	2629	2630	2631	2632	2633	2634	2635	2636	2637	2638	2639	2641	2642
	2643	2644	2645	2646	2647	2648	2649	2650	2651	2652	2653	2654	2655	2656	2657
	2658	2659	2660	2661	2662	2663	2664	2665	2666	2667	2668	2669	2670	2671	2672
	2673	2674	2675	2676	2677	2678	2679	2680	2681	2682	2683	2684	2685	2686	2687
	2688	2689	2690	2691	2692	2693	2694	2695	2696	2697	2698	2699	2700	2701	2702

Cluster No.	Document Number														
	2703	2704	2705	2706	2707	2708	2709	2710	2711	2712	2713	2714	2715	2716	2717
	2718	2719	2720	2721	2722	2723	2724	2725	2726	2727	2728	2729	2730	2731	2732
	2733	2734	2735	2736	2737	2738	2739	2740	2741	2742	2743	2744	2745	2746	2747
	2748	2749	2750	2751	2752	2753	2754	2755	2756	2757	2758	2759	2760	2761	2762
	2763	2764	2765	2766	2767	2768	2769	2770	2771	2772	2773	2774	2775	2776	2777
	2778	2779	2780	2781	2782	2783	2784	2785	2786	2787	2788	2789	2790	2791	2792
	2793	2794	2795	2796	2797	2798	2799	2800	2801	2802	2803	2804	2805	2806	2807
	2808	2809	2810	2811	2812	2813	2814	2815	2816	2817	2818	2819	2820	2821	2822
	2823	2824	2825	2826	2827	2828	2829	2830	2831	2833	2834	2835	2836	2837	2838
	2839	2840	2841	2842	2843	2844	2877	2897	2908	2917	2951	2970	2971	2973	2979
	3001	3005	3006	3059	3075	3112	3137	3156	3159	3181	3198	3225	3256	3262	3264
	3337	3341	3392	3450	3460	3529	3552	3557	3581	3582	3604	3653	3668	3678	3808
	3817	3821	3853												
Cluster #2	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45
	46	47	48	49	50	51	52	53	54	55	56	57	58	59	60
	61	62	63	64	65	66	67	68	69	70	71	72	73	74	75
	76	77	78	79	80	81	82	83	84	85	86	87	88	89	90
	92	93	94	95	96	97	98	100	101	102	103	104	105	106	107
	108	109	110	111	113	114	115	116	117	119	120	121	122	123	124
	125	126	127	128	129	130	131	132	133	134	135	136	137	138	139
	140	141	142	143	144	145	146	147	148	149	150	151	152	153	155
	156	157	158	159	160	161	162	163	164	165	166	167	168	169	170
	171	172	173	174	175	176	177	178	179	180	181	182	183	184	185
	186	187	188	189	190	191	192	193	194	195	196	197	198	199	200
	201	202	203	204	205	206	207	208	209	210	211	212	213	214	215
	216	217	218	219	220	221	222	223	224	225	226	227	228	230	232
	234	235	236	237	238	239	240	241	242	243	244	245	246	247	248
	249	250	251	252	253	254	255	256	257	258	259	260	261	262	263
	264	265	266	267	268	269	270	271	272	273	274	275	276	277	278
	279	280	281	282	283	284	285	286	287	288	289	290	291	292	293
	294	295	296	297	298	299	300	301	302	303	304	305	307	308	309
	310	311	312	313	314	315	316	317	318	319	320	321	322	323	324
	325	326	327	328	329	330	331	332	333	334	336	337	338	339	340
	341	342	343	344	345	346	348	349	350	351	352	353	354	355	356
	357	358	359	360	361	362	363	364	365	366	368	369	370	371	372
	373	374	375	376	377	378	379	380	381	382	383	384	385	386	387
	388	389	390	391	392	393	394	395	396	397	398	399	400	401	402
	403	404	405	406	407	408	409	410	411	412	413	415	416	417	418
	419	420	421	422	423	424	425	426	427	429	430	431	432	433	434
	435	436	437	438	440	441	442	443	444	445	446	447	448	449	450
	451	452	453	454	455	456	457	458	459	460	461	463	464	465	466

Cluster No.	Document Number														
	467	468	469	470	472	473	474	475	476	477	478	479	480	481	482
	483	484	485	486	487	488	489	490	491	493	494	495	496	497	498
	499	500	501	502	503	504	505	506	507	508	509	510	511	512	513
	514	515	516	517	518	520	521	522	523	524	525	526	527	528	529
	530	531	532	533	534	535	536	537	538	539	540	541	542	543	544
	545	546	548	549	550	551	552	553	554	555	556	557	558	559	560
	561	562	563	564	565	566	567	568	569	570	571	572	573	574	575
	576	577	578	579	580	581	582	583	584	585	586	587	588	589	590
	591	592	593	594	595	596	597	598	599	600	601	602	603	604	605
	606	607	608	609	610	611	612	613	614	615	616	617	618	619	620
	621	622	623	624	625	626	627	628	629	630	631	632	633	634	635
	636	637	638	639	640	641	642	643	644	645	646	647	648	649	650
	651	652	653	654	655	656	657	658	659	660	661	662	664	665	666
	667	668	669	670	671	672	673	674	675	676	677	678	679	680	681
	682	683	684	685	686	687	688	689	690	691	692	693	694	695	696
	697	698	699	700	701	702	703	704	705	706	707	708	709	710	711
	712	713	714	715	717	718	719	720	721	722	723	724	725	726	727
	728	729	730	731	732	733	734	735	736	737	738	739	740	741	742
	743	744	745	746	747	749	750	751	752	753	754	755	756	757	758
	759	760	761	762	763	764	765	766	767	768	769	770	771	772	773
	774	775	776	777	778	779	780	781	782	783	784	785	786	787	788
	790	791	792	793	794	795	796	797	799	800	801	802	803	804	805
	806	807	808	809	810	811	812	813	814	815	816	817	818	819	820
	821	822	823	824	825	826	827	828	829	830	831	832	833	834	835
	836	837	838	839	840	841	842	843	844	845	846	847	848	849	850
	851	852	853	854	855	856	857	858	859	860	861	862	863	864	865
	866	867	868	869	870	871	872	873	874	875	876	877	878	879	880
	881	882	883	884	885	886	887	889	890	892	893	894	895	896	897
	898	899	900	901	902	903	904	905	906	907	908	909	911	912	913
	914	915	916	917	918	919	920	921	922	923	924	925	926	927	928
	929	930	931	932	933	934	935	936	937	938	939	940	941	942	943
	944	945	946	947	948	949	950	951	952	953	954	955	956	957	958
	959	960	961	962	963	964	965	966	967	968	969	970	971	972	973
	974	975	977	978	979	980	981	982	983	984	985	986	987	988	989
	990	992	993	994	995	996	997	998	999	1000	1001	1002	1003	1004	1005
	1006	1007	1008	1009	1011	1012	1013	1014	1015	1016	1017	1018	1019	1020	1021
	1022	1023	1024	1026	1028	1029	1030	1031	1032	1033	1034	1035	1036	1038	1039
	1040	1041	1042	1043	1044	1045	1046	1047	1048	1049	1050	1051	1052	1053	1054
	1055	1056	1057	1058	1059	1060	1061	1062	1063	1064	1066	1067	1068	1069	1070
	1071	1072	1073	1074	1075	1076	1077	1078	1079	1080	1081	1083	1085	1086	1087
	1088	1090	1091	1092	1093	1094	1095	1096	1097	1098	1099	1100	1101	1102	1103
	1104	1105	1106	1107	1108	1109	1110	1111	1112	1113	1114	1115	1116	1117	1118
	1119	1120	1121	1122	1123	1124	1125	1126	1127	1128	1129	1130	1131	1132	1133

Cluster No.	Document Number														
	1134	1135	1136	1137	1138	1139	1140	1141	1142	1143	1144	1145	1146	1147	1148
	1149	1151	1152	1153	1154	1155	1156	1157	1158	1159	1160	1161	1162	1163	1165
	1166	1167	1168	1169	1170	1171	1172	1173	1174	1175	1176	1177	1178	1179	1180
	1181	1182	1183	1185	1186	1187	1188	1189	1190	1191	1192	1193	1194	1195	1196
	1197	1198	1199	1200	1201	1202	1203	1204	1205	1206	1207	1208	1209	1210	1211
	1212	1213	1215	1216	1217	1218	1219	1220	1221	1222	1223	1224	1225	1226	1227
	1229	1230	1231	1232	1233	1234	1235	1236	1237	1238	1240	1241	1242	1243	1244
	1245	1246	1247	1248	1250	1251	1252	1253	1254	1255	1256	1257	1258	1259	1260
	1261	1262	1263	1265	1266	1267	1268	1269	1270	1271	1272	1273	1274	1275	1276
	1278	1279	1280	1281	1282	1283	1284	1285	1286	1287	1288	1289	1291	1293	1294
	1295	1296	1298	1299	1301	1303	1304	1305	1306	1307	1308	1309	1310	1311	1312
	1313	1314	1315	1316	1317	1318	1320	1322	1324	1325	1326	1327	1328	1329	1330
	1331	1332	1333	1334	1335	1336	1337	1338	1339	1340	1341	1342	1343	1344	1345
	1346	1347	1348	1349	1350	1351	1352	1353	1354	1355	1356	1357	1358	1359	1360
	1361	1362	1363	1364	1365	1366	1367	1368	1369	1370	1371	1373	1374	1375	1376
	1377	1378	1379	1380	1381	1382	1383	1386	1387	1388	1389	1390	1391	1392	1393
	1394	1395	1396	1398	1399	1400	1401	1402	1403	1404	1405	1406	1407	1408	1409
	1410	1411	1412	1413	1414	1416	1417	1418	1419	1420	1421	1422	1423	1424	1425
	1426	1427	1428	1429	1430	1431	1432	1433	1434	1435	1436	1438	1439	1440	1441
	1442	1443	1444	1445	1446	1492	1576	1813	2094	2163	2190	2413	2546	2832	2874
	3153	3195	3202	3448	3449	3453	3457	3459	3462	3472	3562	3563	3652	3731	3732
	3733	3734	3739	3824											
Cluster #3	91	99	112	118	154	231	414	428	471	748	798	888	891	910	976
	1025	1065	1089	1164	1184	1264	1290	1292	1297	1300	1302	1321	1437	1589	1933
	2061	2161	2403	2640	2845	2846	2847	2848	2849	2850	2851	2852	2853	2854	2855
	2856	2857	2858	2859	2860	2861	2862	2863	2864	2865	2866	2867	2868	2869	2870
	2871	2872	2873	2875	2876	2878	2879	2880	2881	2882	2883	2884	2885	2886	2887
	2888	2889	2890	2891	2892	2893	2894	2895	2896	2898	2899	2900	2901	2902	2903
	2904	2905	2906	2907	2909	2910	2911	2912	2913	2914	2915	2916	2918	2919	2920
	2921	2922	2923	2924	2925	2926	2927	2928	2929	2930	2931	2932	2933	2934	2935
	2936	2937	2938	2939	2940	2941	2942	2943	2944	2945	2946	2947	2948	2949	2950
	2952	2953	2954	2955	2956	2957	2958	2959	2960	2961	2962	2963	2964	2965	2966
	2967	2968	2969	2972	2974	2975	2976	2977	2978	2980	2981	2982	2983	2984	2985
	2986	2987	2988	2989	2990	2991	2992	2993	2994	2995	2996	2997	2998	2999	3000
	3002	3003	3004	3007	3008	3009	3010	3011	3012	3013	3014	3015	3016	3017	3018
	3019	3020	3021	3022	3023	3024	3025	3026	3027	3028	3029	3030	3031	3032	3033
	3034	3035	3036	3037	3038	3039	3040	3041	3042	3043	3044	3045	3046	3047	3048
	3049	3050	3051	3052	3053	3054	3055	3056	3057	3058	3060	3061	3062	3063	3064
	3065	3066	3067	3068	3069	3070	3071	3072	3073	3074	3076	3077	3078	3079	3080
	3081	3082	3083	3084	3085	3086	3087	3088	3089	3090	3091	3092	3093	3094	3095
	3096	3097	3098	3099	3100	3101	3102	3103	3104	3105	3106	3107	3108	3109	3110
	3111	3113	3114	3115	3116	3117	3118	3119	3120	3121	3122	3123	3124	3125	3126
	3127	3128	3129	3130	3131	3132	3133	3134	3135	3136	3138	3139	3140	3141	3142

Cluster No.	Document Number														
	3143	3144	3145	3146	3147	3148	3149	3150	3151	3152	3154	3155	3157	3158	3160
	3161	3162	3163	3164	3165	3166	3167	3168	3169	3170	3171	3172	3173	3174	3175
	3176	3177	3178	3179	3180	3182	3183	3184	3185	3186	3187	3188	3189	3190	3191
	3192	3193	3194	3196	3197	3199	3200	3201	3203	3204	3205	3206	3207	3208	3209
	3210	3211	3212	3213	3214	3215	3216	3217	3218	3219	3220	3221	3222	3223	3224
	3226	3227	3228	3229	3230	3231	3232	3233	3234	3235	3236	3237	3238	3239	3240
	3241	3242	3243	3244	3245	3246	3247	3248	3249	3250	3251	3252	3253	3254	3255
	3257	3258	3259	3260	3261	3263	3265	3266	3267	3268	3269	3270	3271	3272	3273
	3274	3275	3276	3277	3278	3279	3280	3281	3282	3283	3284	3285	3286	3287	3288
	3289	3290	3291	3292	3293	3294	3295	3296	3297	3298	3299	3300	3301	3302	3303
	3304	3305	3306	3307	3308	3309	3310	3311	3312	3313	3314	3315	3316	3317	3318
	3319	3320	3321	3322	3323	3324	3325	3326	3327	3328	3329	3330	3331	3332	3333
	3334	3335	3336	3338	3339	3340	3342	3343	3344	3345	3346	3347	3348	3349	3350
	3351	3352	3353	3354	3355	3356	3357	3358	3359	3360	3361	3362	3363	3364	3365
	3366	3367	3368	3369	3370	3371	3372	3373	3374	3375	3376	3377	3378	3379	3380
	3381	3382	3383	3384	3385	3386	3387	3388	3389	3390	3391	3393	3394	3395	3396
	3397	3398	3399	3400	3401	3402	3403	3404	3405	3406	3407	3408	3409	3410	3411
	3412	3413	3414	3415	3416	3417	3418	3419	3420	3421	3422	3423	3424	3425	3426
	3427	3428	3429	3430	3431	3432	3433	3434	3435	3436	3437	3438	3439	3440	3441
	3442	3443	3444	3445	3446	3447	3451	3452	3454	3455	3456	3458	3461	3463	3464
	3465	3466	3467	3468	3469	3470	3471	3473	3474	3475	3476	3477	3478	3479	3480
	3481	3482	3483	3484	3485	3486	3487	3488	3489	3490	3491	3492	3493	3494	3495
	3496	3497	3498	3499	3500	3501	3502	3503	3504	3505	3506	3507	3508	3509	3510
	3511	3512	3513	3514	3515	3516	3517	3518	3519	3520	3521	3522	3523	3524	3525
	3526	3527	3528	3530	3531	3532	3533	3534	3535	3536	3537	3538	3539	3540	3541
	3542	3543	3544	3545	3546	3547	3548	3549	3550	3551	3553	3554	3555	3556	3558
	3559	3560	3561	3564	3565	3566	3567	3568	3569	3570	3571	3572	3573	3574	3575
	3576	3577	3578	3579	3580	3583	3584	3585	3586	3587	3588	3589	3590	3591	3592
	3593	3594	3595	3596	3597	3598	3599	3600	3601	3602	3603	3605	3606	3607	3608
	3609	3610	3611	3612	3613	3614	3615	3616	3617	3618	3619	3620	3621	3622	3623
	3624	3625	3626	3627	3628	3629	3630	3631	3632	3633	3634	3635	3636	3637	3638
	3639	3640	3641	3642	3643	3644	3645	3646	3647	3648	3649	3650	3651	3654	3655
	3656	3657	3658	3659	3660	3661	3662	3663	3664	3665	3666	3667	3669	3670	3671
	3672	3673	3674	3675	3676	3677	3679	3680	3681	3682	3683	3684	3685	3686	3687
	3688	3689	3690	3691	3692	3693	3694	3695	3696	3697	3698	3699	3700	3701	3702
	3703	3704	3705	3706	3707	3708	3709	3710	3711	3712	3713	3714	3715	3716	3717
	3718	3719	3720	3721	3722	3723	3724	3725	3726	3727	3728	3729	3730	3735	3736
	3737	3738	3740	3741	3742	3743	3744	3745	3746	3747	3748	3749	3750	3751	3752
	3753	3754	3755	3756	3757	3758	3759	3760	3761	3762	3763	3764	3765	3766	3767
	3768	3769	3770	3771	3772	3773	3774	3775	3776	3777	3778	3779	3780	3781	3782
	3783	3784	3785	3786	3787	3788	3789	3790	3791	3792	3793	3794	3795	3796	3797
	3798	3799	3800	3801	3802	3803	3804	3805	3806	3807	3809	3810	3811	3812	3813
	3814	3815	3816	3818	3819	3820	3822	3823	3825	3826	3827	3828	3829	3830	3831

Cluster No.	Document Number														
	3832	3833	3834	3835	3836	3837	3838	3839	3840	3841	3842	3843	3844	3845	3846
	3847	3848	3849	3850	3851	3852	3854	3855	3856	3857	3858	3859	3860	3861	3862
	3863	3864	3865	3866	3867	3868	3869	3870	3871	3872	3873	3874	3875		

Data set ‘YahooK1’(misclassified document)

Cluster No.	Document Number														
Cluster # 1	495	496	497	498	499	500	502	504	505	507	508	510	511	517	518
	519	525	526	534	539	544	545	546	547	548	555	556	558	559	561
	562	564	565	567	576	577	578	580	581	582	583	584	585	586	588
	591	606	609	610	611	612	613	614	615	616	617	618	619	620	621
	626	627	629	630	631	632	633	634	635	636	644	647	648	649	650
	651	653	664	665	666	669	677	678	681	683	689	690	699	703	710
	711	712	713	714	715	716	719	720	721	722	723	724	737	738	741
	742	743	745	752	753	757	758	762	763	765	766	767	770	771	772
	776	789	790	794	795	796	797	798	799	800	801	803	804	807	808
	811	812	813	814	815	816	824	825	827	828	829	830	831	832	834
	835	836	837	838	839	840	841	844	848	851	857	862	866	868	869
	870	871	877	878	879	880	881	882	883	884	885	886	887	889	893
	894	898	913	934	935	936	937	942	943	944	945	946	947	948	949
	952	956	957	958	959	960	961	962	963	965	966	967	968	973	976
	977	978	979	986	987	995	996	1002	1003	1011	1012	1015	1016	1017	1018
	1021	1026	1027	1028	1029	1030	1039	1040	1041	1042	1048	1049	1057	1058	1059
	1060	1062	1063	1064	1065	1067	1068	1069	1070	1071	1072	1073	1074	1075	1076
	1077	1079	1080	1084	1086	1087	1089	1090	1091	1094	1099	1100	1101	1102	1104
	1105	1106	1107	1108	1109	1110	1112	1121	1122	1123	1124	1132	1136	1139	1142
	1143	1148	1149	1150	1151	1152	1153	1171	1174	1175	1176	1177	1178	1179	1180
	1181	1182	1183	1184	1186	1187	1189	1193	1198	1201	1209	1211	1212	1213	1214
	1215	1219	1220	1228	1229	1230	1231	1233	1234	1235	1236	1237	1238	1268	1269
	1270	1271	1272	1273	1274	1275	1278	1279	1280	1281	1282	1283	1284	1293	1294
	1298	1302	1306	1311	1312	1313	1314	1315	1316	1317	1318	1319	1320	1328	1331
	1333	1335	1336	1337	1338	1342	1345	1349	1350	1351	1352	1353	1354	1355	1362
	1364	1366	1367	1368	1369	1372	1376	1394	1397	1398	1399	1400	1401	1402	1403
	1404	1405	1409	1435	1440	1441	1442	1445	1449	1460	1461	1462	1468	1470	1471
	1472	1473	1474	1475	1476	1477	1478	1479	1486	1487	1488	1490	1491	1492	1495
	1497	1512	1513	1514	1517	1518	1521	1522	1523	1524	1525	1527	1528	1530	1531
	1536	1537	1538	1543	1544	1546	1547	1548	1549	1553	1554	1556	1557	1559	1570
	1571	1572	1575	1576	1577	1578	1579	1580	1581	1583	1584	1585	1586	1593	1594
	1595	1596	1611	1614	1615	1616	1617	1621	1623	1624	1631	1632	1633	1634	1635
	1636	1637	1638	1639	1642	1643	1644	1645	1646	1647	1648	1649	1650	1651	1654
	1655	1657	1658	1659	1660	1661	1662	1663	1664	1669	1676	1681	1687	1689	1690

Cluster No.	Document Number														
	1691	1692	1694	1695	1696	1697	1698	1699	1700	1701	1704	1705	1710	1713	1737
	1738	1739	1742	1743	1744	1745	1746	1747	1748	1749	1750	1751	1752	1753	1754
	1755	1756	1766	1767	1770	1781	1783	1784	1785	1786	1793	1799	1805	1806	1807
	1820	1821	1822	1823	1825	1838	1840	1841	1842	1843	1849	1850	1851	1852	1855
	1857	1859	1860	1866	1867	1868	1869	1870	1875	1880	1883	2203	2208		
Cluster # 2	1	2	3	4	5	6	7	8	9	10	12	13	14	15	16
	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32
	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48
	49	50	51	52	53	54	55	56	58	59	60	61	62	63	64
	65	66	67	68	69	70	71	72	73	74	75	77	78	79	80
	81	82	83	84	85	86	87	88	89	90	91	92	93	94	95
	96	97	98	101	102	103	104	105	106	107	108	109	110	111	112
	113	114	115	118	119	120	121	122	124	126	127	128	129	130	131
	132	133	134	135	136	137	138	139	140	141	142	143	144	145	146
	147	148	150	151	152	153	155	156	157	158	159	160	161	162	163
	164	165	166	168	169	170	171	172	173	174	175	176	177	178	179
	180	181	184	185	186	187	188	189	190	191	192	193	194	195	196
	197	198	199	200	201	202	203	204	205	206	207	208	209	210	211
	212	213	214	215	216	217	218	219	220	221	222	223	224	225	226
	229	230	231	232	233	234	235	236	237	238	239	240	241	242	243
	244	245	246	247	248	249	250	251	252	253	254	255	257	258	259
	260	261	262	263	264	265	266	267	268	269	270	272	273	274	275
	276	277	278	279	280	281	282	283	284	285	286	287	288	289	290
	291	292	293	294	295	296	297	298	299	301	302	303	304	305	306
	307	308	309	311	312	313	314	315	316	317	318	319	320	321	322
	323	324	325	326	327	328	329	330	331	332	333	334	336	337	338
	339	340	341	342	343	344	346	347	348	349	350	351	352	353	354
	355	356	357	358	359	360	361	362	363	364	365	366	367	368	369
	370	371	372	373	374	375	376	377	378	379	380	381	382	383	384
	385	386	387	388	389	390	391	392	393	394	395	396	397	398	399
	401	402	403	405	406	407	408	409	410	411	412	413	414	415	416
	417	418	419	420	421	422	423	424	425	426	427	428	429	430	431
	432	433	434	435	436	437	438	439	441	442	443	444	445	446	447
	448	449	450	451	452	453	454	455	457	458	459	460	461	462	463
	464	466	467	468	469	470	471	472	473	475	476	477	478	479	480
	481	482	483	484	485	486	487	488	489	490	491	492	493	494	506
	744	852	1303	1332	1656	1732	2309	2335							
Cluster #3	76	99	100	116	117	123	125	149	154	167	182	183	228	256	271
	310	335	345	400	404	465	474	528	536	543	560	563	602	608	623
	624	637	638	668	675	676	692	693	694	701	717	718	725	736	739
	740	746	748	749	751	755	759	781	785	791	802	818	842	843	845
	846	849	850	856	858	859	860	861	888	895	896	900	902	903	904
	914	938	950	953	969	970	991	998	1019	1036	1037	1044	1046	1144	1146

Cluster No.	Document Number														
	1147	1163	1164	1165	1190	1191	1192	1207	1208	1216	1217	1240	1241	1242	1256
	1257	1258	1285	1288	1289	1290	1291	1295	1296	1300	1305	1309	1343	1344	1359
	1370	1371	1373	1375	1410	1411	1412	1413	1414	1415	1416	1417	1418	1419	1421
	1423	1424	1425	1426	1427	1428	1433	1434	1436	1437	1438	1446	1447	1448	1469
	1482	1483	1484	1493	1494	1504	1509	1515	1529	1539	1540	1551	1587	1588	1589
	1590	1597	1600	1601	1608	1609	1613	1622	1627	1628	1630	1652	1653	1665	1671
	1672	1673	1674	1677	1678	1679	1680	1682	1693	1702	1703	1706	1707	1714	1715
	1716	1717	1718	1719	1720	1721	1722	1723	1724	1725	1726	1733	1734	1735	1740
	1741	1768	1769	1771	1772	1773	1774	1775	1776	1777	1778	1779	1780	1787	1788
	1790	1795	1796	1797	1800	1801	1802	1803	1804	1808	1809	1810	1811	1824	1826
	1829	1830	1831	1832	1833	1834	1835	1836	1848	1853	1854	1856	1871	1872	1873
	1876	1878	1879	1881	1882	2207									
Cluster #4	11	17	33	57	503	512	529	533	537	538	540	541	542	553	557
	575	579	589	640	643	672	686	688	695	702	730	768	786	853	874
	897	912	915	924	925	926	927	928	929	930	931	932	954	964	1043
	1066	1083	1166	1169	1196	1197	1204	1248	1250	1255	1276	1277	1287	1301	1365
	1374	1444	1481	1498	1500	1503	1550	1558	1736	1798	1874	1884	1885	1886	1887
	1888	1889	1890	1891	1892	1893	1894	1895	1896	1897	1898	1899	1900	1901	1902
	1903	1904	1905	1906	1907	1908	1909	1910	1911	1912	1913	1914	1915	1916	1917
	1918	1919	1920	1921	1922	1923	1924	1925	1926	1927	1928	1929	1930	1931	1932
	1933	1934	1935	1936	1937	1938	1939	1940	1941	1942	1943	1944	1945	1946	1947
	1948	1949	1950	1951	1952	1953	1954	1955	1956	1957	1958	1959	1960	1961	1962
	1963	1964	1965	1966	1967	1968	1969	1970	1971	1972	1973	1974	1975	1976	1977
	1978	1979	1980	1981	1982	1983	1984	1985	1986	1987	1988	1989	1990	1991	1992
	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007
	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022
	2023	2024	2025	2026	2027	2028	2030	2031	2032	2033	2034	2035	2036	2037	2038
	2039	2040	2041	2042	2043	2044	2045	2046	2047	2048	2049	2050	2051	2052	2053
	2054	2055	2056	2057	2058	2059	2060	2061	2062	2063	2064	2065	2066	2067	2068
	2069	2070	2071	2073	2074	2075	2076	2077	2078	2079	2080	2081	2082	2083	2084
	2085	2086	2087	2088	2089	2090	2091	2092	2093	2094	2095	2096	2097	2098	2099
	2100	2101	2102	2103	2104	2105	2106	2107	2108	2109	2110	2111	2112	2113	2114
	2115	2116	2117	2118	2119	2120	2121	2122	2123	2124	2125	2126	2127	2128	2129
	2130	2131	2132	2133	2134	2135	2136	2137	2138	2217	2235	2236	2237	2238	2261
	2308														
Cluster #5	227	300	440	456	522	530	594	595	596	598	622	642	670	682	687
	700	727	728	729	731	732	733	734	735	747	754	778	784	787	788
	833	847	855	890	892	901	906	974	975	992	999	1000	1001	1038	1052
	1098	1130	1141	1145	1188	1199	1200	1202	1227	1243	1244	1246	1247	1249	1253
	1292	1297	1304	1307	1308	1341	1443	1499	1501	1507	1582	1591	1592	1598	1602
	1603	1607	1612	1668	1675	1728	1729	1730	1791	1792	1877	2029	2072	2139	2140
	2141	2142	2143	2144	2145	2146	2147	2148	2149	2150	2151	2152	2153	2154	2155
	2156	2157	2158	2159	2160	2161	2162	2163	2164	2165	2166	2167	2168	2169	2170

Cluster No.	Document Number														
	2171	2172	2173	2174	2175	2176	2177	2178	2179	2180	2181	2182	2183	2184	2185
	2186	2187	2188	2189	2190	2191	2192	2193	2194	2195	2196	2197	2198	2199	2200
	2201	2202	2204	2205	2206	2209	2210	2211	2212	2213	2214	2215	2216	2218	2219
	2220	2221	2222	2223	2224	2225	2226	2227	2228	2229	2230	2231	2232	2233	2234
	2239	2240	2241	2242	2243	2244	2245	2246	2247	2248	2249	2250	2251	2252	2253
	2254	2255	2256	2257	2258	2259	2260	2262	2263	2264	2265	2266	2267	2268	2269
	2270	2271	2272	2273	2274	2275	2276	2277	2278	2279	2280	2281	2282	2283	2284
	2285	2286	2287	2288	2289	2290	2291	2292	2293	2294	2295	2296	2297	2298	2299
	2300	2301	2302	2303	2304	2305	2306	2307	2310	2311	2312	2313	2314	2315	2316
	2317	2318	2319	2320	2321	2322	2323	2324	2325	2326	2327	2328	2329	2330	2331
	2332	2333	2334	2336	2337	2338	2339	2340							
Cluster # 6	501	509	513	514	515	516	520	521	523	524	527	531	532	535	549
	550	551	552	554	566	568	569	570	571	572	573	574	587	590	592
	593	597	599	600	601	603	604	605	607	625	628	639	641	645	646
	652	654	655	656	657	658	659	660	661	662	663	667	671	673	674
	679	680	684	685	691	696	697	698	704	705	706	707	708	709	726
	750	756	760	761	764	769	773	774	775	777	779	780	782	783	792
	793	805	806	809	810	817	819	820	821	822	823	826	854	863	864
	865	867	872	873	875	876	891	899	905	907	908	909	910	911	916
	917	918	919	920	921	922	923	933	939	940	941	951	955	971	972
	980	981	982	983	984	985	988	989	990	993	994	997	1004	1005	1006
	1007	1008	1009	1010	1013	1014	1020	1022	1023	1024	1025	1031	1032	1033	1034
	1035	1045	1047	1050	1051	1053	1054	1055	1056	1061	1078	1081	1082	1085	1088
	1092	1093	1095	1096	1097	1103	1111	1113	1114	1115	1116	1117	1118	1119	1120
	1125	1126	1127	1128	1129	1131	1133	1134	1135	1137	1138	1140	1154	1155	1156
	1157	1158	1159	1160	1161	1162	1167	1168	1170	1172	1173	1185	1194	1195	1203
	1205	1206	1210	1218	1221	1222	1223	1224	1225	1226	1232	1239	1245	1251	1252
	1254	1259	1260	1261	1262	1263	1264	1265	1266	1267	1286	1299	1310	1321	1322
	1323	1324	1325	1326	1327	1329	1330	1334	1339	1340	1346	1347	1348	1356	1357
	1358	1360	1361	1363	1377	1378	1379	1380	1381	1382	1383	1384	1385	1386	1387
	1388	1389	1390	1391	1392	1393	1395	1396	1406	1407	1408	1420	1422	1429	1430
	1431	1432	1439	1450	1451	1452	1453	1454	1455	1456	1457	1458	1459	1463	1464
	1465	1466	1467	1480	1485	1489	1496	1502	1505	1506	1508	1510	1511	1516	1519
	1520	1526	1532	1533	1534	1535	1541	1542	1545	1552	1555	1560	1561	1562	1563
	1564	1565	1566	1567	1568	1569	1573	1574	1599	1604	1605	1606	1610	1618	1619
	1620	1625	1626	1629	1640	1641	1666	1667	1670	1683	1684	1685	1686	1688	1708
	1709	1711	1712	1727	1731	1757	1758	1759	1760	1761	1762	1763	1764	1765	1782
	1789	1794	1812	1813	1814	1815	1816	1817	1818	1819	1827	1828	1837	1839	1844
	1845	1846	1847	1858	1861	1862	1863	1864	1865						