ROUGH SET BASED CLUSTERING FOR FINDING RELEVANT DOCUMENT

NG CHOON CHING

UNIVERSITI MALAYSIA PAHANG

ROUGH SET BASED CLUSTERING FOR FINDING RELEVANT DOCUMENT

NG CHOON CHING

THESIS SUBMITTED IN FULLFILLMENT OF THE DEGREE OF COMPUTER SCIENCE (SOFTWARE ENGINEERING)

FACULTY OF COMPUTER SYSTEM & SOFTWARE ENGINEERING

UNIVERSITI MALAYSIA PAHANG

2013

ABSTRACT

Searching for relevant documents based on the keywords of particular selected articles are proposed in this thesis. This method is proposed to help user get relevant document based on the articles they selected. The common searching engine will return up to thousand articles where some articles are not really relevant to the searching too. In this paper, rough set-based data mining technique is employed to enhance the result of searching relevant documents. The rough set-based clustering technique, namely Min-Min Roughness (MMR) is applied to cluster documents from Wikipedia into groups according to keywords of selected articles in the effort for finding relevant documents. This research is done using dataset of articles from online Wikipedia website. The proposed keywords methods for finding relevant documents will save time during searching progress. This research is expected to be useful for finding relevant documents.

ABSTRAK

Idea untuk mencari dokumen yang berkaitan berdasarkan kata kunci tertentu artikel-artikel terpilih telah dicadangkan dalam tesis ini. Kaedah ini dicadangkan untuk membantu pengguna mendapatkan dokumen yang berkaitan berdasarkan rencana yang mereka telah memilih. Enjin carian biasa akan bagi sehingga berribu artikel di mana beberapa artikel tidak benar-benar berkaitan dengan pencarian juga dicadang. Dalam kertas ini, teknik perlombongan data set berasaskan kasar digunakan untuk meningkatkan hasil daripada mencari dokumen yang berkaitan. Kasar teknik kelompok berasaskan set, iaitu Kekasaran Min-Min (MMR) digunakan untuk dokumen kelompok dari Wikipedia ke dalam kumpulan mengikut kata kunci artikel terpilih dalam usaha untuk mencari dokumen yang berkaitan. Kajian ini dilakukan dengan menggunakan dataset artikel dari laman web Wikipedia talian. Cadangan kaedah kata kunci untuk mencari dokumen yang berkaitan akan menjimatkan masa dalam mencari kemajuan. Kajian ini dijangka akan berguna untuk mencari dokumen yang berkaitan.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER I

INTRODUCTION

Firstly, this chapter is briefly discussing about the overview of research. This chapter includes only six parts. First part of this chapter is background; followed by the problem statement. Then the following part is the motivation and after this are the scopes of this research. The fifth part of this chapter is the objectives which means determine the research's goal. Last part is the thesis organization where the structure of the thesis is briefly described.

1.1. Background

Knowledge Discovery from Databases (KDD) is the non-ordinary extraction of formerly unknown, hidden, and potentially favorable information from the data. KDD includes some amount of automated methods which can mine useful information from the data stored in the databases as a branch of machine learning.

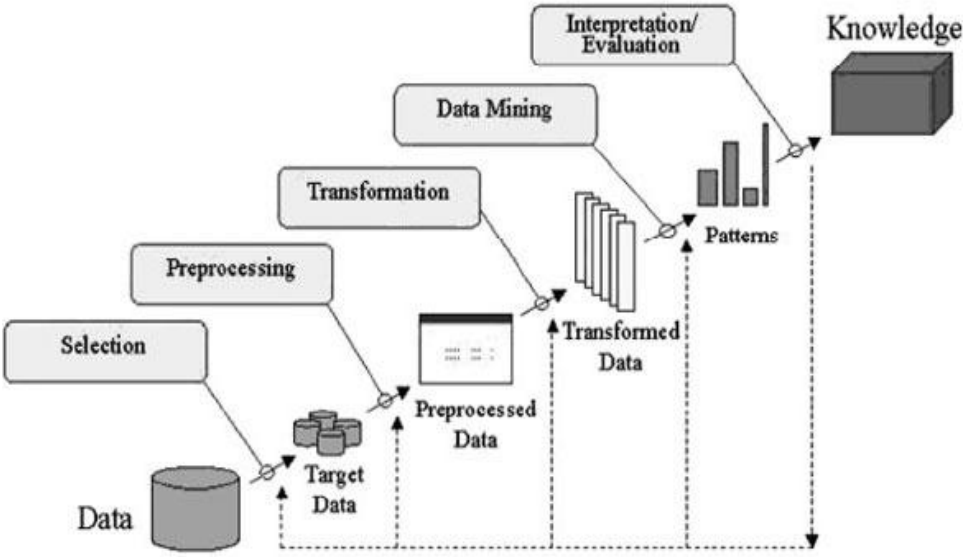

Figure 1.1: KDD Process [1]

Figure 1.1 shows the KDD process. There are many steps in KDD process. One of the steps is data mining which apply algorithms of data analysis and discovery during acceptable limitations of computational efficiency, manage to produce a specifically enumeration of the data's pattern. Among the methods of data mining includes clustering and classification. Clustering describe the data by identifying a finite set of clusters or categories, whereas classification involving learning function that classifies or maps a data item onto one of the several predefined classes.

Document cluster means gather all similar documents together into a group which the documents similar in the certain constraints so that relevant documents can be grouped together. The main aim to cluster the document is to locate the natural groupings; hence the overview of the main points or topics in a group of documents can be presented and seen clearly. For example, document clustering is applied in the official website portal of the United State government to automatically form the search results into relevant classes automatically. When the word "immigration" is submitted, the results a user can see as the relevant categories such as "Citizenship and Immigration Services", "Employment", "Department of Homeland Security", "Immigration Reform" and others. There is a different between clustering and classification [2]. For classification the properties of a group are already known then only the related documents are put into

that group whereas for clustering the all available documents are grouped based on the similarity criteria such as keywords. There are many techniques can be used to cluster the documents in effort to find the relevant document. In general, document cluster (or text cluster) is considered as parts of the data clustering which including some concept of NLP (natural language processing), ML (machine learning) and also IR (information retrieval). There are many algorithms can be applied in the techniques of document clustering. In this research, rough set are applied to find relevant documents based on the important keywords from particular articles selected. In other words, when a particular document or articles are selected, the important keywords of tat articles are detected and then used as the attributes in rough set theory to search for other articles which are relevant to it.

1.2. Problem Statement and Motivation

To date, the number of articles or documents which available in Wikipedia or other online sources are getting more and more. Almost everyday there are new articles or documents are added to different online sources including Wikipedia. The current search engine and recommendation available will normally returns up to thousands of documents when the user hit the search button to search a word, this bring difficulties for users to detect and browse which is the relevant information they want. It is really very time consuming if user chooses to browse all the thousands pages of articles. So, the demand of user to have a better searching result of documents which is really relevant to what they are trying to find. Currently the famous search engine such as in google will return lots of documents which including those not really closely relevant to what users searching for. Moreover, when the user has a really nice article and they want to get another document which are relevant to the current article they own now, there are still do not have any suitable tools or application which manage to do so. So, the use of rough set theory and data mining can solve the problem. This will surely provide more useful and relevant articles or documents for the user's searching result. Thus, user can save time and achieve better searching result to a greater extend.

1.3. Objectives

The objective of this study is mainly to enhance and evaluating results for searching relevant documents. There are four objectives in this research:

a. To do research of finding relevant documents via rough set theory
b. To cluster the documents which are related to each other into a meaningful group based on the keywords of selected article.
c. To give a suitable recommendation of articles to users to improve the results of searching.

1.4. Scopes

The scopes of this research are described as follow.

a. The dataset used are created from the Online Wikipedia articles.
b. The clustering uses rough set-based clustering technique.
c. The keywords are used as the attribute for MMR computation to find relevant documents.

1.5. Thesis Organization

The organization of the thesis is described as follows. Chapter II describes the idea and concept of KDD and document clustering. Chapter III explains about rough set theory, modeling process, dataset and min-min roughness data clustering. Chapter IV discusses the results of research in apply rough set theory for clustering data to find relevant documents and following by discussion. Lastly, Chapter V describes the conclusion of this work.

CHAPTER II

LITERATURE REVIEW

First, currently existing literatures which concerning with the proposed research is briefly explained in this chapter. This chapter is divided into three sections, which is the simple information about Knowledge Discovery in Databases (KDD), general idea of Data Mining (DM) and some information regarding document clustering.

2.1. Knowledge Discovery in Databases (KDD)

In this section, the definition, processes, examples and applications of Knowledge Discovery in Databases (KDD) are presented.

2.1.1. Definition of KDD

The field of knowledge discovery in databases (KDD) is getting to be very famous and has grown rapidly recently. To date, all data are being collected and accumulated at a high speed due to the large variety of domain. According to Frawley et al. (1992), the quantities of information in our world are predicted to get twice as much for every twenty months [3]. Besides, Fayyed et al. (1996) found that there is a critical demand for a new era of computational theories and tools to support human beings in extracting beneficial information or knowledge from the rapidly increase volumes of digital data [4]. These principles and instruments are the main part of the emerging domain of KDD. KDD is defined as the process of determining relevant, original, potentially useful, and finally clear and simple patterns in data. Moreover, Carter et al. defined KDD is

thenontrivial extraction of implicit, previously unknown, and potentially beneficial information from data [5]. As a component of machine learning, KDD includes a number of automated methods whereby useful information is mined from data stored in databases. It is a must to have efficiency in handling the huge input datasets which usually can be found in commercial surrounding while implementing the KDD method as a functional tool for discovering knowledge from databases. The performance of three KDD algorithms which are similar to each other being brought together and compared to identify its applicability to the huge commercial databases.

According to Zhong et al. KDD is usually a multiple phase of process which involve abundant steps such as preprocessing, data preparation, search for hypothesis generation, pattern formation, knowledge evaluation, representation, refinement and management[6]. All those processes are carried out to perform the systems which involve the KDD techniques efficiently. Furthermore, the data in databases are usually being updated frequently so that the information stored in databases is always up to date. Therefore, at different phases or levels, the process may be iterated (Fayyad et al., 1996) [7].

The main objective of KDD is to illustrate the patterns of data which humans can understand easily. M.S.Chen claimed that KDD is a process of nontrivial extraction of reserved, formerly unknown and potentially useful information (such as knowledge rules, constraints, regularities) from data in databases [8]. Those extracted pattern are beneficial for the users to do their tasks. KDD enable the interesting knowledge, regularities, or high-level information to be extracted from different angles, and huge databases thereby serve as rich and trustworthy sources for confirmation and generation.

2.1.2. KDD Processes

There are generally some steps in KDD process which consists of the usage of database together with whichever needed selection, preprocessing, subsampling and transformations of it; applying data mining methods or algorithms to list the patterns from it; and assessing the products of data mining to determine the subset of the assessed patterns deemed knowledge. Besides, the assessment and probable

interpretation of the mined patterns which used to identify which patterns is most suitable to be new knowledge also included in KDD process.

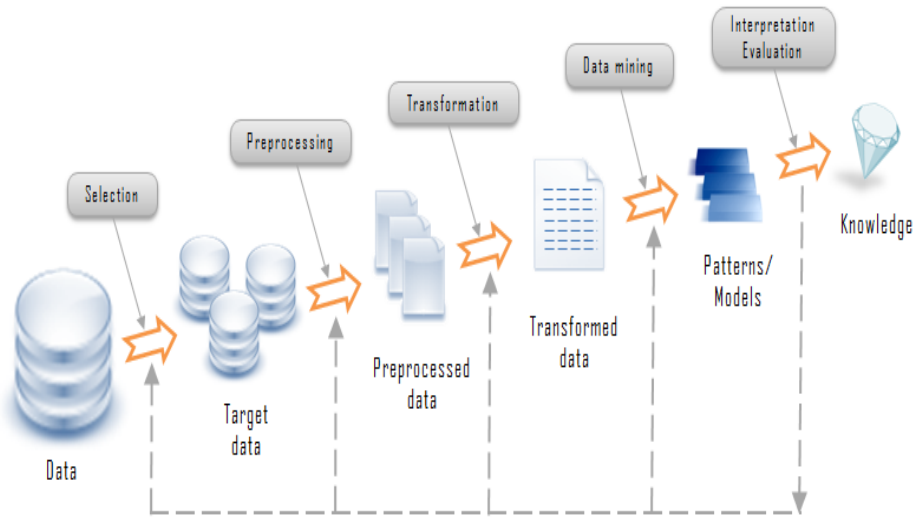

Figure 2.1: The overview of steps that build up the KDD process [9]

From Figure 2.1, there are some steps in the process of KDD such as the below:

a. Exploit an understanding of the application domain, goal(s) of end user, and the relevant prior knowledge. This initial step of process will equip the user to understand and develop the objectives of the application thoroughly.
b. Create or select the target data set. The user need to select or establish what data should be used, or focus on a subset of variables or data samples, on which discovery is to be performed.
c. Data cleaning and preprocessing. This step involves some operations such as remove the noise and outliers. Essential information is also collected to model or account for noise. It also acts as strategies for handling missing data, and accounting for time-sequence information and known changes.
d. Data reducing and projection. In this step, useful features are found to indicate the data depending in objective of the task. It may consist of methods such as reduction and transformation to minimize the effective number of variables under consideration or to detect the invariant representations of the data.
e. Choose data mining task. The objectives of KDD process is determined if it is summarization, classification, regression, clustering, etc.

f. Select algorithms of data mining. In order to search for useful patterns from dataset, selecting method(s) is used. In this process, judgment regarding which parameters or models might be more suitable is made. The specific of data mining method is matched with the entire criteria of KDD process.
g. Data mining. Data mining algorithms including classification rules or trees, regression, and clustering are implemented to search interested patterns in a specific representational form.
h. Interpret mined patterns. This step comprises the evaluation and the interpretation of the discovered patterns. This step has possibility of going backwards to any of steps 1 to 7 for further usage. For example, some features can be added to step 5 and steps are repeated again from there.
i. Use the discovered knowledge. In this final step, the knowledge is incorporated into another system for further action such as changes may be made to the system and effect of changes are measured.

2.1.3. Examples of KDD

There are some examples of application which involve KDD methods. One example of KDD application areas in marketing is database marketing systems which analyze customer databases to evaluate different customer group and predict their activity. For example, American Express (AMEX) had reported that there are 10 - 15 percent increases in the card usage with the aid of KDD method. Another famous marketing application is market-basket analysis systems which able to determine the patterns such as "If customer bought x, he/she is also likely to buy Y and Z." These interesting patterns are precious to retailers and can help them to know more details about their investment. Besides, KDD is also applied at investment areas. Quite a number of companies' use data mining in the investment field but most of them do not describe clearly their system. But, LBS Capital Management is one of the exceptions. Its system uses expert system, neural nets, and genetic algorithms to manage portfolios which total up to $600 million, since the system begin in 1993 it has outperformed the board stock market. Furthermore, fraud detection is also included in KDD application area. For example in the effort to monitor the credit card trick, HNC Falcon and Nestor PRISM

systems are used to observe over the millions of accounts. At the same time, the FAIS system from the U.S. Treasury Financial Crimes Enforcement Network is also used wisely to determine the financial transactions that may show money laundering activity.
Moreover, manufacturing is also one of the KDD application areas. The ASSIOPEE troubleshooting system is developed as part of a joint venture between General Electric and SNECMA, was applied by three major European airlines to diagnose and predict problems for the Boeing 737. This system used clustering methods to derive the families of faults. While in telecommunications, the TASA (telecommunications alarm-sequence analyzer) was built in cooperation with a manufacturer of telecommunications equipment and three telephone networks. The novel framework is used by system to locate frequently occurring alarm episodes from the alarm stream and present them as rules. The system also manages to explore large set of discovered rules together with flexible information-retrieval tools supporting interactivity and iteration. In this way, TASA offers grouping, pruning, and ordering tools to refine the results of basic brute-force search for rules.

2.1.4. Application of KDD in computer science fields

Generally, computer science means the study of a branch of science which deal with computer no matter is hardware or software. It is true that KDD had contributed and had a lot of potential in computer science field especially in science data analysis. As the data and information are getting more and more, so database are created to store all those data. When the database are huge and many, it is hard for human being to find useful information that are needed. Hence KDD application plays an important role in computer science field too.
Recently the vase amount of data is due to the widely usage of bar codes for government transactions, hypermarket, and some business which are computerized. As usual users store all these data in computer database so that they able to access it easily. The sudden expansion of database usage leads to the urgent need of technology or technique such as KDD to help extract useful information from large databases. Researches have combined KDD together with computer science knowledge, together with algorithms to develop systems which manage to extract valuable data. The

development of KDD in computer science field is clearly shown that being led by the demand of users want to extract and transform large dataset into valuable knowledge efficiently.

2.2. Data Mining

In this section, the definition, examples and applications of Data Mining (DM) are presented.

2.2.1. Definition of DM

Fayyed et al. define data mining as a process of extracting and identifying useful information using techniques such as statistical, mathematical, artificial intelligence and machine learning, which knowledge can be gained from large database [10]. Generally, society always define data mining similar as KDD, or many other term for instance information discovery, information harvesting, knowledge extraction, data archeology, data pattern processing, mining knowledge from huge databases, data analysis and etc. which is defined as to find or extract the favorable and interesting patterns of data. Nevertheless, KDD is a field of computer science, which includes the tools and theories which help in extracting useful and previously unknown information from large collection of digitized data, and KDD contains several steps whereas data mining is one of the steps in KDD. DM is an application of a particular algorithm which used in extracting the patterns from data.

Besides, data mining is the essential factor in KDD. A new idea to organize and manage the huge data is hence provided. Data mining is also being explained as an interdisciplinary field with an ordinary objective to predict the outcomes and separate the complex relationships in data. Moreover, the duplicated iterative application of specific methods of data mining is sometimes get involved in the data mining component of the KDD process. Automated tools involving complex algorithms are used in data mining to find out hidden patterns, associations, anomalies or structure from big amounts of data which is keep information repositories or data storehouse.

There are many types of data mining tasks such as discover new pattern to describe data, predict the characteristics of pattern and many more. Generally, the three components of combination which make up the data mining algorithm are as follow:

a. Model representation: there are descriptions about the patterns discovered. The two corresponding factors are model's function (such as classification and clustering) and its representational form (such as linear function of multiple variables and a Gaussian probability density function).
b. Criteria of model evaluation: the basis criteria or preference of model such as how well a model or pattern meets the objectives (fit function) is important. For example the novelty, accuracy and many more is the criteria need to be noted during the implementation data mining algorithms.
c. Search method: search the specific parameters of algorithms to maximize the usage of particular algorithms or given model representation.

Data mining tasks are divided into two category which are prediction methods (classification, regression, time series analysis, prediction) and descriptive methods (clustering, summarization, association rules, and sequence discovery). Predictions methods mean use a few variables to predict the unknown value of other variables whereas a descriptive method is to find the understandable pattern which clearly describe the data.

a. Classification: Data are mapped into predefined classes or groups. As the classes are determined before the examination of data, it is always being referred to as supervised learning.
b. Regression: A data item is mapped to an actual valued prediction variable. In regression, the target data is assumed to be fitted into some function and then the best function of this model from given data is determined.
c. Time series analysis: As the value varies over time it is examined. Normally the value picked has evenly spaced of time points. During time series analysis, the time series plot is used for visualization.
d. Prediction: Can be seen as one type of classification. The future states are predicted based on the current and past data instead of predicting the current state.
e. Clustering: data item is mapped to one of the clusters. It is alternatively refer to unsupervised learning as well as segmentation.

f. Summarization: data is mapped into subsets with combination of simple explanation. It is also known as generalization or characterization.
g. Association rules: To identify the particular type of data associations. This refers to the data mining task of exposing the relationships between data.
h. Sequence discovery: Used to identify the sequential pattern from dataset based on the activity in time series.

2.2.2. Examples of DM

Data mining applications play important roles in many fields such as finance, marketing, telecommunications and many more. For example in the finance field, some company will issue financial reports which hid real status of the company, then investor would not know the real financial status of that company. This would make investors don't know the financial distress of the company they are investing hence bring lost to investors. So, financial prediction system or model should be constructed to help investors and company. Data mining technique are included in developing the system so that can extract useful variable of companies' financial status. Besides, in marketing field, marketing strategies are close related with the information of transaction and customer in the databases. Attractive pattern of customer information and the way they buy things and shopping will be hidden without the adequate tools to analyze the data from huge databases. Suitable data mining techniques are used to tackle unknown information from large database. For example, customer profiles are important to be tackled as it can be used to meet the needs of customers and reduce the mailing cost. Besides, data mining can help to detect the customer favor based on the data from transaction database. Moreover, data mining also play a role in insurance sector. Nowadays society would buy insurance no matter for personnel, car, house and many more. There are many packages which amount of claim is different. Hence, an application using data mining is needed to help customer to predict the claims which more suitable for them based on the criteria such as accident rates and age of the car for car insurance. With the systems using data mining technique can help customer to sign the insurance plan which suit them the most. There is another example of application which use data mining technique is DBMiner. DBMiner is actually a data mining

system. A big range of data mining functions such as clustering, association, classification; comparison, prediction and characterization are implemented in DBMiner system. The great amount of data from data warehouse and relational database are being analyzed via DBMiner. In short, DBMiner can help users to get excellent results during extracting useful information in shortest time.

2.2.3. Applications of DM in computer science fields

To date, the information and amount of huge database in real life keep on increasing from time to time especially in computer science field. It is a big project to record down all the information and analyzes data in large database. Hence DM methods are used to extract the useful data from that entire large database. There are some examples of application of DM in science field such as geologist use data mining method in helping them to find the volcanoes on Venus. The databases of images and digital libraries are getting bigger from time to time, that is the reason data mining systems which able to find content are used here. Besides, data mining play an important role in biological sequence database. Nowadays, the structures as well as the function of molecules are determined using data mining methods. The application of DM can make the process to extract information become simple and save time. Sky survey cataloging is also applying data mining methods in it such as classification. The images are segmented and attributes of image (features) are measured as well. Then based on those features, the class is modeled. This enable scientist to make prediction of the class (galaxy or star) of sky objects especially those objects can't be seen easily by eyes, based on the images get from telescopic (from Palomar Observatory).

2.3 Document Clustering

In this section, the definition, preprocessing dataset and common techniques of document clustering are presented.

2.3.1 Definition of Document Clustering

Since decade ago, it is already clearly seen that human being already have the idea to sort similar or relevant things into categories or groups, for instance, human already and be able to recognize that lots of individual objects actually sharing some properties such as this group of foods are edible or poisonous. Brian Everitt et al. defines clustering as when an amount of objects are given, each of them are described by a set of numerical measures, devise a classification scheme for grouping the objects into a number of classes such that the objects inside a class are similar in some respect and is different from other classes. Moreover, the number of classes and special characteristics of each single class are to be determined [11]. Among all the unsupervised learning clustering can be said as the most common form of it. Karger et al. claims that society not really well receive the usage of document clustering compared to the common information retrieving tool due to some reasons [12]. The main reason people not prefer to use document clustering compared to information retrieval tools is the running time for clustering is too slow working on a huge corpora and large amount of documents. Besides, people feel that clustering actually does not obviously improve the retrieval but then this problem only occur when use clustering in the attempt to enhance the common search techniques. However, the preprocessing that done on the dataset before start the clustering using rough set theory may enhance the performance of retrieval. Clustering is often wrongly referred to as an automatic classification which is incorrect, this is because clusters found are not recognize former to processing whereas in the classification cases, the classes all are already pre-defined. In clustering, there is the allocation and the property of data that will determine the membership of cluster, in opposition to the classification where the classifier learns the association between objects and classes from training set, which is a set of data that are labeled accurately by manual, and then reproduce the learning behavior n unlabeled data [13].

2.3.2 Preprocessing

Before start to cluster the documents, the preprocess steps of the text in documents need to be done [14]. During preprocessing, all the terms which are not contain important

content will be removed, meanwhile, the ending words also being removed to let it be the original word. Below are some examples of steps involved in the preprocessing stage:

a. Filter: the whole text of the documents need to be filtered to remove the unnecessary things such as the punctuations, citations, common words (is, are, the), special characters which not play any important role in document clustering. If the text is taken form a web pages then the colors or pictures of the background should totally being exclude too.
b. Tokenization: the whole full sentences are now being split into single individual words. All the same words will be put together.
c. Stemming: this is the process to get the basic form of all the words in document. For instance, the basic form of these words "relates", "related", "relation" will be "relate". Nowadays there are many algorithms such as porter stemmer can be used to do the stemming process.
d. Stopword removal: a stopword means a term that is not considered to be the important criteria to cluster the document. For example if the words are repeated then it is "removed" in the sense that the word is put aside first since it is appeared earlier in the document already.
e. Pruning: the words which appear lesser time is considered low frequency words and will be removed too. There is an assumption such as these low frequency words are not very useful in clustering even the words have whichever differentiating power. But there are also some special cases which the words appear too frequent such as more than 40% in the documents are also being removed.

2.3.3 Techniques in Document Clustering

K-Means Clustering

K-means clustering (MacQueen, 1967) [15] is one of the methods or techniques being used for clustering. It is considered as a simple but yet a famous algorithms used in clustering as it can produce a particular amount of disjoint and the flat (also known as non-hierarchical) clusters. This method is considered as numerical, iterative, not deterministic and also unsupervised. By using this method, all the important objects

involved had to be indicated as a collection of numerical characteristics. The amount of groups (which called as k) that one wants to identify can be specified via the usage of k-means. There are some common properties for the k-means algorithms such as K clusters are present and in each of the cluster there is always have minimum one item in it. Besides, the clusters so not overlap with each other therefore it is non-hierarchical. As the "center" of clusters is not involved with the closeness, hence each item of a cluster can be stay closer with its own cluster rather than any other cluster.

The simple flow of k-means algorithms are as follow:

a. Partition the obtained dataset into K clusters and randomly allocate the data points to the clusters so that the clusters can probably have the similar amount of data point.
b. The distance from the data point to every single cluster is computed for every data points obtained.
c. Determine whether if the data point is nearest to its own cluster, if it does then just leave it there else shift the data points into the nearest cluster.
d. Steps above are iterated until get the results which do not have any data points shifting from one cluster to another cluster when go through all the data points. At this stage the clusters can be said to be stable and can end the process of clustering.
e. Take note that the selection of primary partition may extremely influence the final outcomes, which in term of inter-cluster and of course intracluster distances and cohesion as well.

However, there are some pros and cons for using the k-means method to do clustering. The main advantage of using k-means value is when the K is small, the k-means could be calculated faster compare to the hierarchical clustering with a large amount if variables. K-means also can generate closer cluster compared with the hierarchical clustering specifically if the clusters are globular. On the other hand, using k-means methods will let user face some trouble when compare the quality of the clusters generated such as for distinct K value or primary partition which may affect the clustering results. Furthermore, k-means not really work perfectly with those non-globular clusters and the fixed amount of cluster can cause the difficulty in predicting the value of K should actually be.