IMPROVING THE ACCURACY OF TEXT DOCUMENT CLUSTERING BASED ON SYNGRAM ALGORITHM

ABDUL HALIM BIN OMAR

A thesis submitted in
fulfillment of the requirement for the award of the
Degree of Master of Information Technology

Faculty of Computer Science and Information Technology
Universiti Tun Hussein Onn Malaysia

SEPTEMBER, 2015

ABSTRACT

In most of the literature, Vector Space Model (VSM) represents text document by the frequencies of terms occurred inside the document. In general, the relationship between terms that appear in text document has been ignored by VSM. As a result, two major limitations of term relationship are treated as single and independent entities. The limitation of both concepts, such as Polysemy and Synonymy are definitely significant in determining the content of text document. To overcome both limitations, this study has proposed a combination of WordNet and N-grams named as Syngram algorithm. WordNet is selected as a lookup database to obtain synonym concepts. The capabilities of both concepts are introduced to overcome the Synonymy limitation in text documents into sequences of synonym sets. In the second approach, N-grams have been used in language modeler to construct the term consecutive. This study exploited N-grams to defy Polysemy limitation by altering text features into chunks of terms. The transformation of frequent single term to frequent concept has been proven to improve the accuracy of the text document clustering. An experiment was conducted on reuters50_50 dataset with 10 classes of author names and the performance is compared with existing algorithms. The experiment results showed that the proposed algorithm (65.6%) outperformed the existing algorithm VSM (55.4%), N-grams (53.2%) and WordNet (59%).

ABSTRAK

Sorotan kajian terdahulu telah banyak menyatakan penggunaan *Vector Space Model (VSM)* sebagai satu kaedah bagi mewakilkan teks dokumen. Perwakilan itu dilakukan dengan mengambil kira kekerapan istilah-istilah yang telah wujud di dalam teks dokumen. Secara umumnya, Istilah-istilah tersebut telah diabaikan hubungan diantara mereka dan *VSM* menukarkan istilah-istilah tersebut kepada suatu entiti yang tunggal. Oleh kerana itu, *VSM* telah mengakibatkan dua permasalahan utama yang berpunca daripada pengabaian tersebut. Kedua-dua permasalahan ini merupakan *Polysemy* dan *Synonymy* konsep. Bagi mengatasinya, kajian ini telah mencadangkan pengabungan dua kaedah iaitu *WordNet* dan *N-grams* yang dinamakan sebagai algorithma Syngram. *WordNet* telah dipilih kerana ia merupakan sebuah pangkalan data yang dapat memberikan konsep-konsep sinonim yang mana dapat digabungkan supaya menjadi urutan set sinonim. Kaedah kedua ialah *N-grams*, ia merupakan suatu kaedah kebarangkalian yang telah digunakan dalam pemodelan bahasa dan ia cukup bermanfaat bagi menghasilkan urutan-urutan istilah. Oleh yang demikian, kajian ini telah mengeksploitasikan *N-grams* dalam menyelesaikan masalah *Polysemy* dengan mengubah istilah-istilah ke dalam bentuk ketulan urutan istilah. Dengan mengubah istilah teks dokumen dari kekerapan tunggal kepada kekerapan berkonsep, ia terbukti telah meningkatkan prestasi (ketepatan) text document clustering. Satu eksperimen telah dijalankan ke atas *reuters50_50* dataset dengan 10 kelas nama pengarang dan hasil eksperimen telah dibandingkan antara text document clustering *(k-means)* dengan *VSM, N-grams, WordNet* dan algoritma yang telah dicadangkan. Keputusan telah menunjukkan algorithma cadangan (65.6%) telah mengatasi VSM (55.4%), N-gram (53.2%) dan WordNet (59%).

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS AND ABBREVIATIONS

w_i	Weighting of text document
tf_i	Term Frequency
idf_i	Inverse Document Frequency
df_i	Document Frequency
D, d	Document
log()	Log Formulation for Normalization
Synsets_i	Set of synonym
C_n	Concept of term
sf_i	Syngram Frequency
Dictionary	WordNet Dictionary
Synsets_n	Set of synonym
Syn	Synonym term
Term_1	Term
$\text{Term}_i\|\text{Syn}_{i;}$	Concatenation between terms
$P(c_1^n)$	Probability of concept
x_1	Variable x_1
∩	Interception
$recall$	Subject in F-Measure
F	F-Measure
P	Precision
tn	True negative
fn	False negative
tp	True Positive
fp	False Positive
R	Recall
F-Score	Combination score of Precision and Recall

$\cos(d_i, d_j)$	Cosine angle between documents
TF-IDF	Term Frequency Inverse Document Frequency
VSM	Vector Space Model
BOW	Bag of Word
WordNet	Electronically lexical Dictionary
N-grams	Probability of Chain Rules
Synonymy	Concept of Synonym
Polysemy	Concept of Polysemy (Same term but has multiple meaning)
F- Measure	Measurement Formulation for Clustering
HTML	Hyper Text Markup Language
XML	Extensible Markup Language
DOM	Document Object Model
K-means	Partitioning algorithm for clustering
Syngram	Proposed algorithm (Synsets + N-grams)
Reuters50_50	Text documents dataset from UCIMLR
NLP	Natural Language Processing
Apriori	Apriori Algorithm
TF-IDF	Term Frequency Inverse Document Frequency
CCAT	Class Criteria Cognitive Aptitude Test
Synset	Set of Synonym from WordNet
Euclidean	Method to Calculate Distance
JAVA	Programming Language
UCIMLR	University California Irvine Machine Learning Repository
Hierarchical	Hierarchical Clustering Algorithm
Spiral	Spiral Clustering Algorithm
DBSCAN	DBSCAN Clustering Algorithm
Synonyms	Same Meaning of terms
Hypernyms	Hierarchical of Categorize Terms
Hyponyms	Anatomy of Term
Meronyms	Explaining Hyponyms
VPNs	Virtual Private Network Security
IDS	Intrusion Detecting System

KUSZA	Sultan Zainal Abidin Religious College
UTHM	University Tun Hussein Onn Malaysia
KUiTTHO	University College Tun Hussein Onn Malaysia

LIST OF APPENDICES

CHAPTER 1

INTRODUCTION

1.1 An Overview

Internet is known as a participative medium has been designed for the whole world. Via the internet, user would be able to broadcast any ideas or running any services over the internet by utilizing website as a primary platform. Basically website is a regular medium for public to carry out their network activities such as social networking, doing business transaction, as a learning occasion and many more. Those mentioned activities requiring data and the data processed into information that might be considered either useful or harmful to everyone. Toby, Collind & Jammy (2009) have defined some collective of data might be presented in many forms, either it is unstructured, semi-structured and structured. Initially, data presented in a free form or arbitrary sizes and types. However, several frameworks such as Hyper Text Markup Language (HTML) and Extensible Markup Language (XML) invented to encapsulate data in semi structured form known as Document Object Model (DOM). The finest form is structured data which is stated in the specific location in database. The database stores the data in precise and complete formatted. This formatted scheme has ensured the data to become more significant and efficient to be managed. On the other hand, the data lies inside document and processed into the nature of information that kindly being used nowadays.

Basically document has come in many sizes and forms such as images, texts, sounds and videos. The highest amount of information available online was formed in a textual documentation by indicating approximately 80% of document over the

internet are stored in the form of text (Yu Xiao, 2010). It is consistent with rapidly growing of the internet user within this information age, the information spread from side to side through the websites and it turns out to be overloaded which brought a lot of choices of information. These choices of information have made text document become a good sources of references.

Despite of all the positive outcomes of having it as a good source of information, text document still remains unstructured and needs to be clustered into a significant and more meaningful collection. Moreover, a lot of researches have been done with regards to cluster the text document, which refers to structure unstructured text document in a huge set of corpus and concentrating on text clustering algorithm. As a result, there are many text clustering algorithms existed over industry and some of well-known algorithms are K-means, Hierarchical, Spiral Model and many more. The purpose of the listed clustering algorithm is to solve the issues related to the unstructured text document.

Text document clustering algorithm can be defined as a task of separating text documents into homogeneous classes or clusters into their own related groups. In the process of separating text documents, the text documents in mutual classes must be same as possible while text documents in the contradicting classes must be dissimilar as possible. In Figure 1.1 shows clustered text documents as objects represented by blue, green and red color. That conceptual figuration is a collection of object grouped together based on similar color. Those objects have been connected by edges in order to show the distances between every object that being clustered.

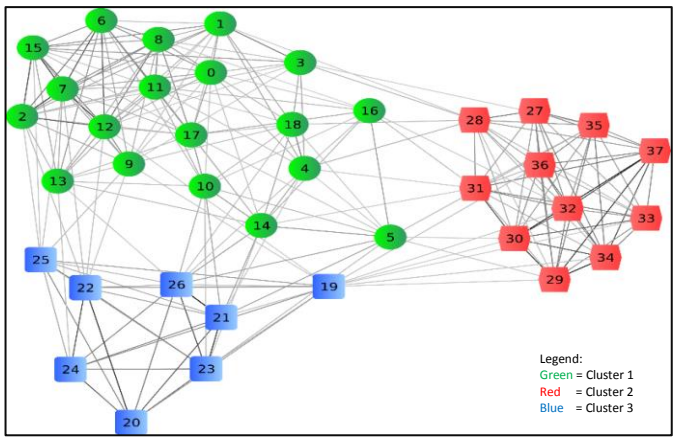

Figure 1.1: Group of Clustered object

There are many types of text clustering algorithm existed, but the most popular are hierarchical and partitioning. Nevertheless, both algorithms have shared a mutual objective which is to cluster the text documents. Before clustering any text documents, one important thing to be considered as a step before clustering process is text document conversion. It is become crucial since the text document clustering is working on numerical data and requires an outstanding method to convert text documents into numerical values.

The conversion of the text documents into numerical values is unavoidable routine before deploying any text document clustering. It has been started in early 1975, a professor of Computer Science at Cornell University has founded the Vector Space Model (VSM). This technique was successfully applied in information retrieval (Salton, Wong & Yang, 1975). It is very useful and widely being used in text document conversion.

Basically, text document conversion based on VSM similarly works as Bag of Words (BOW). It treats all term occurred in all text documents by independently, which mean term will be separated into single terms (Baghel & Dhir, 2010). With this approach it has suffered a limitation regarding term relationship which is very important in measuring text document similarity. To further justify, the reason of the VSM limitation is due to the existence of relationships occurred inside the text documents. Furthermore, VSM only emphasizes on single term instead of hidden term concepts that important to be digested.

The terms represent the content or idea written inside text document that was determined by the authors. In some occasions the terms appeared in text document share a same form but may have a different meaning. This phenomenon can be stated as Polysemy concept. However, the terms appeared in the text documents may share the same meaning in a different form, called as the Synonymy concept. Both issues are extremely related to VSM since it only looks into the frequent single term and overlooked the term relationship inside the text documents. In a clear view of the term relationship, a linguistic researcher Petho (2001) has revealed a concept that lies inside a group of terms as Polysemy and Synonymy concepts. Both are much related to this research as they are considered reality concepts that took place in text documents. Petho (2001) has located the meaning of Polysemy as a phenomenon when a single term has multiple meanings. This means that the Polysemy concept may express different things in different contexts. This concept can be confusing in

VSM, as example of “drives me crazy” and “driving a car”. They may look simple to be understood due to different forms and contexts, but in VSM, “drive me crazy” and “driving a car” are the same.

In contrast, the concept of Synonymy is about terms that have the same meaning but appear in different forms like “big house” and “huge house”. Both term concepts are so important and it is compulsory to take them into account before applying text clustering algorithm in order to increase the quality of text clustering result. Figure 1.2 is a simple version of the Polysemy concept by Vaquero, Saenz & Barco (2000) and Vaquero *et al.,* 2000 mentioned that Term 2 has two meanings and it's important in determination of meaning 1 and meaning 2.

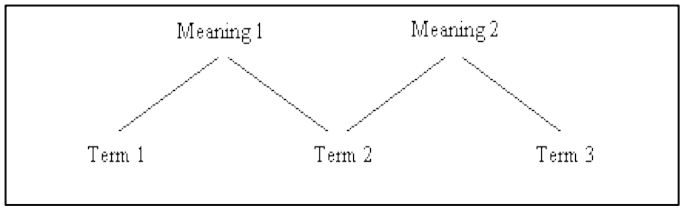

Figure 1.2: Term Concept of Polysemy

Moreover, Terms 3 and Term 1 are different to each other but in the meaning determination, they shared the same connector, which is Term 2. This connector is more likely to have multiple meanings yet reflect same form. Also, the connection between terms will be constructed into phrase that has a longer meaning rather than the one with single term.

In conclusion, the term relationship or term concept is important to be concerned instead of being ignored by VSM. This ignorant will cause lack of the accuracy to text document clustering while it performing. It is very important since VSM is the main key of the process before calculating the distance between the text documents.

1.2 Problem Statements

The big issue in text document clustering is about text document similarity. In order to determine the similarity, text documents must be converted into numerical values to make sure clustering algorithm capable to compute the similarities. Therefore, VSM is one of the popular technique being used to convert the text documents into sequences of numerical values. Unfortunately, VSM cause two major problems which are Polysemy and Synonymy concepts. Both concepts are really important in determining the accuracy of text document clustering since most of text document clustering depends on VSM (Baghel & Dhir, 2010). The VSM works on counting the frequent single term and ignoring the term concept. This behavior is extremely make a distance of text documents become indistinct and not properly measured. In literature, the used of term concept is very useful to counter the Synonymy problem instead of original single term. Thus, the VSM might be improved by changing the text features and retrieving term synonym concept from a lexical dictionary. Many scholars (Huang *et al.*, 2008; Hamou *et al.*, 2010; Ray & Singh, 2010; Thanh & Yamada, 2011; Bouras & Tsogkas, 2012; Celik & Gungor, 2013) used WordNet as a platform to extract term relationship to apply on research in finding Synonymy concept. The Synonym concept has proven improved the clustering performance in term of accuracy because all synonym term are concatenated and become singleton. On the other hand, the Polysemy concept is about term order. It's indicated by retrieving the term set of text documents will make VSM become more understandable and possible to fix the quality of clustering result. N-grams are one of the methods to generate term consecutive by using chain rule. This chain rule are really beneficial, Alneyadi & Muthukkumarasamy (2013) used N-grams chain rule to generate consecutive term inside text documents. The consecutive term is purposely to distinguish the pair of terms in text content analysis studies. Besides that, WordNet and N-grams is possible to be combined. Recently research has done by Go & See (2008), which is combining both methods to solve the text dimensionality problem. The problem addressed by Go & See (2008) is arisen due to the problem caused by N-grams without considering the accuracy of text document clustering. As a result, WordNet and N-grams was combined and producing frequent concept of N-grams consecutive that reduce the text document dimensionality. Buscaldi *et al.* 2012 also combined WordNet and N-grams. The combination was to study the differences

between conceptual or semantic similarity of text fragments in text documents by utilizing N-grams as method to detect term frequencies. Both researchers combine WordNet and N-gram based on their own preferences. In this study WordNet and N-grams being utilized is to increase the text document clustering accuracy. The challenge of this research is to prove the combination of WordNet and N-grams have the significant effects after changing the text features from frequent term into frequent concept in improving text document clustering algorithm (K-Means).

1.3 Objectives of the Study

The objectives of this research:

i. To propose Syngram algorithm based on a frequent concept of Polysemy and Synonymy.
ii. To improve text document clustering by deploying the proposed algorithm
iii. To evaluate the performance of the proposed algorithm based on accuracy of text document clustering result by using F-Measure.

1.4 Scope of the Study

This research is focusing on the improvement of text document clustering (K-means) accuracy by utilizing the frequent concepts instead of frequent single term. The performances of proposed algorithm (frequent concept) and the existing algorithm (frequent term) were compared and analyzed in order to identify which approach is better. A common dataset reuters50_50 obtained from University California Irvine Machine Learning Repository (UCIMLR) was used as a sample for experimental process. The experiment were carried out by using RapidMiner 5.0 on Pentium i5 with 3.0 GHz Acer Workstation, 8.0 GB RAM and focusing on English text document.

1.5 Aim of the Study

The aim of the study is to improve the result of text document clustering (K-means) base on VSM (frequent single term). The improvement was made by deploying Syngram algorithm (frequent concept).

1.6 Significance of the Study

This study investigated the performances of text document clustering with frequent term and frequent concept. It was discovered in this study that text document clustering with frequent concept improved further accuracy compared to frequent term. The frequent concept was originally constructed from the combination of WordNet lexical dictionary and N-gram chain rule. This combination has improved the performance of text document clustering instead of frequent term approach that was implemented by previous researchers.

1.7 Outline of the Thesis

This thesis consisted of five chapters, including Chapter one. Following is the summarization of each chapter.

(i) ***Chapter 1*****:** *Introduction*. Apart from providing an outline of the thesis, this chapter contains an overview of the research background, problem statement, objectives, scope, aim, and significance of the study.

(ii) ***Chapter 2*****:** *Literature Review*. This chapter included a review on VSM limitation regarding Polysemy and Synonymy concepts and also reviews the term relationship in WordNet dictionary which cover the Synonymy concept and N-grams chain rule for Polysemy concept. Furthermore a review on previous researches regarding clustering text documents with frequent concept which is involving WordNet, N-grams and incorporation of WordNet and N-grams was also been done. The last review is about K-means clustering algorithm which chosen for data testing in this research. At the end of this chapter, some of the advantages of using WordNet and N-grams are outlined. This chapter lays a foundation for introducing a

new method in improving the text document clustering accuracy by proposing the algorithm as described in Chapter 3.

(iii) ***Chapter 3*****:** *Research Methodology*. This chapter discusses the research methodology used to carry out the study systematically. Initially started with dataset selection, pre-processing, propose algorithm, applying clustering algorithm and cluster evaluation. The main subject in this chapter is regarding a new algorithm called as Syngram has been proposed. The Syngram will further explained on how it worked in order to improve the accuracy of text document clustering.

(iv) ***Chapter 4*****:** *Result and discussion*. The proposed algorithm in Chapter 3 is further validated for its accuracy improvement in this chapter. The performances of the proposed algorithm were tested for comparison against the conventional VSM, WordNet and N-grams. The performance evaluation was carried out based on the document clustering quality of mapping cluster label and with precision and recall.

(v) ***Chapter 5*****:** *Conclusions and Future Works*. The contributions of the proposed algorithm are summarized and the recommendations are described for further continuation of work.

CHAPTER 2

LITERATURE REVIEW

2.1 Introduction

This chapter provides the literature review for better understanding on related issues of Vector Space Model (VSM). In this chapter, an investigation of VSM is conducted intentionally to reveal the limitations and the current improvement of the VSM regarding term conceptual or term relationship. It is related to the issues that associate with text document similarity since the VSM is frequently being used for text document conversion before deploying any text clustering algorithm. This study also reviews on VSM as important role to determine the similarity between the text documents and by concentrating on the limitations of VSM some improvements can be proposed to enhance the performance of text document clustering. VSM has two major limitations which are really significant in determining the content of text documents. These two major limitations are about different term share the same meaning (Synonymy) and term contact that share the same term meaning by constructed phrase (Polysemy). Both limitations are treated as major because by ignoring these Synonymy and Polysemy concept the content of text documents will become misinterpreted. This chapter explains the recent works on text document clustering with term concept. In this Chapter also has discovered the research gap or finding which the contribution of this study.

2.2 Text Document Clustering

Text document clustering is one of the methods used in many data mining application. It manage to group text documents based on their similarity criterion. In a text document clustering the group of similar documents must be more similar between intra documents and less similarity between intra documents of two clusters (Elahi & Rostami, 2012). As a result the similarity of text document can be measured by the formulation of the distance.

There are many algorithms in text document clustering have been developed to cluster the text documents. The clustering algorithm was available in different type of approaches. Many scholars (Zhao & Karypis, 2005; Zhou & Yu, 2011; Suyal, Panwar & Singh, 2014) done the research regarding text document clustering generally to pursue the automatic indexing of document retrieved and it based on similarity characteristic of the text document. It is important to know that text document clustering sometimes is often confused with text classification because of their owned characteristics in classifying the object. But in reality, both algorithms have two dissimilarities, in which text classification requires a predefined label to predict the pattern as opposed to text clustering while text clustering does not require a training set of data (Ravichandra, 2003). Furthermore, the standard of text clustering algorithm is usually separated into two groups, namely partitioning algorithm and hierarchical algorithm (Bharati & M. Ramageri, 2010). In general process, the hierarchical clustering agglomerates the object by visiting one by one of the object. Once the object similarity meets the requirement, the hierarchical of similar object will be constructed. While the partition clustering works on dividing the object into partition and calculate the mean of every object that close to the centroid which is been allocated. Both have owned advantages and it is proven in the comparative study of some common text document clustering techniques (Steinbach, Karypis & Kumar, 2000). In particular, the comparison between two main approaches involves the agglomerative hierarchical clustering (hierarchical clustering) and K-means (partitioning clustering). As a result partitioning clustering performed better than hierarchical clustering in term of response time and the accuracy is belong to hierarchical.

Table 2.1 shows several types of the partitioning clustering algorithms and hierarchical clustering algorithm. Both types of algorithms are categorized based on the respective characteristics or traits when the text document is clustered.

Table 2.1: Partitioning and Hierarchical Clustering

Partitioning Clustering Algorithm	Hierarchical Clustering Algorithm
K-means	Single Linkage
X-Means	Average Linkage
Bisecting K-means	Complete Linkage

There are four basic steps to be implemented in deploying any text document clustering as shown in in Figure 2.1.

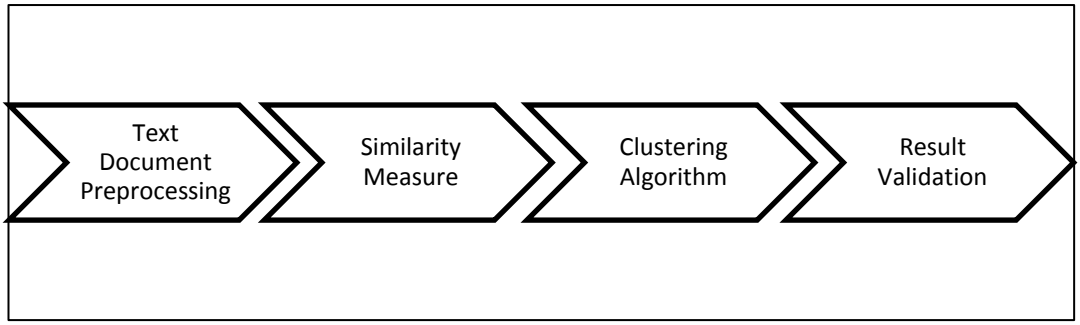

Figure 2.1: Basic Steps in Text Document Clustering (Ravichandra, 2003)

The first step is referred as text document preprocessing, where it considered as a basic and an essential process to take place before applying text document clustering. The second step is similarity measure. This step responsible to consider the similarity between text documents and it has come in assortment of methods, one of the most frequently being used in text document clustering is cosine similarity (Kalaivendhan & Sumaithi, 2014). The cosine method is used to determine the text documents similarity.

The third step in text document clustering is selecting a clustering algorithm. This step uses a particular similarity measure as a subroutine process. This step choose the best algorithm to cluster the text document because by selecting the right clustering algorithm, more or less would be able to get the performance increased

(Ravichandra, 2003). Furthermore inside clustering algorithm subroutines, the distance of text documents must be calculated in order to distinguish the text documents to the centroid point. This centroid point is calculated by determining the mean values of every text documents vector. Based on the calculated centroid point, clustering algorithm will determined the document cluster by grouping them into mutual clusters regarding the distances of text documents to the closest centroid point.

Most of clustering algorithms quantify the distance between a point and the cluster centroid for separating them into relevant groups. It is a vital process in order to create a mutual cluster among the text documents (Ravichandra, 2003). After mutual clusters have been setup, the final process in text document clustering procedure is result validation which is generating the output in statistical figure with the number of clustered text document. It is the last step in validating the clustering result and this last step is the process of iteration in clustering at the prior stage, which is having some methods depend on the clustering task of validation.

Ravichandra (2003) mentioned that the steps in a text document clustering involve calculating the similarity between text documents before selecting the clustering algorithm. In order to calculate the similarity, it is compulsory to convert text documents into sequences of numerical values to make sure text documents would be compared. This conversion is known as a weighting process, which also can be classified as a preprocessing stage of text document and it is important for the smoothness and accuracy of the clustering algorithm.

2.3 Text Document Preprocessing

In text document clustering, data preprocessing is considered as an important stage before deploying any text clustering algorithm. It consisted of five processes, which lie solely on the content of text document. Data preprocessing also considered as a sub process of every text clustering outline and it is important since text documents content various superfluous symbol or character that influence the superiority of clustering performance. Figure 2.2 has illustrated five stages of process in a purification of text document by purging unwanted punctuations and characters.

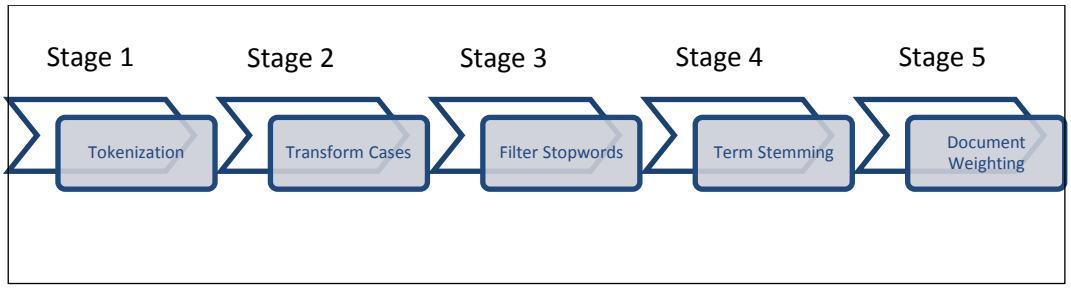

Figure 2.2: Five Stages of Data Preprocessing Text Documents

This study only emphasizing on Document Weighting since VSM worked in this stage. In Document Weighting process, VSM works to convert the text document into sequences of numerical values and every term inside text document will be separated independently. Based on the separation of terms into singleton it cause the ignorant of term relationship (Polysemy and Synonymy concept) which is really important to text document clustering. The ignorant of term relationship will caused a major mistake in distinguishing the text documents. This chapter will explained more detail regarding text document weighting and the limitations of Polysemy and Synonymy concepts.

2.3.1 Text Document Weighting

Text document weighting is a conversion process of text documents into vectors that are represented in numerical (Srividhya & Anitha, 2011). This conversion will allows text documents to be measured by changing the text document features into sequences of numerical values. This conversion method is very significant for text document clustering in order to determine the object distances while clustering the text document.

Most of text clustering algorithm used VSM as basic conversion of text document into numerical vector values (Beil, Ester & Xu, 2002). The vector values will be assigned by length of vector and applying length normalization in order to be compared among text documents. After the conversion process has been done, the next process is to compare the similarity between text documents. In this process, require some methods such as cosine, jaccard, dice, manhattan and many more. Once

the similarity method has been chosen, the distances between text documents could be calculated and compared successfully.

2.3.2 Text Document Conversion with VSM

Vector Space Model (VSM) is basically inspired from an algebraic model in representation of text document and generally it can be practiced on any object to count the frequencies (Salton *et al.*, 1975). It is been widely used in many appliance such in information retrieval, information filtering, information indexing and so on. VSM is also often used in text clustering in utilizing Natural Language Processing (NLP) since text document was an object that represents human in expressing something.

Furthermore in VSM, terms inside text document is treated as a Bag of Word (BOW). This BOW was established by consuming term as features from the collection of text documents. By the implementation of this model it will corresponded to the text documents as a collection of terms by ignoring all term relationship and word order (Baghel & Dhir, 2010). Nevertheless, VSM is frequently being used in text classification and text clustering because this method does takes into account the frequent term occurred inside text document. It is like a bunch of frequent term extracted from a collection of text document. In order to understand this term representation, Table 2.2 below shows the process of text document conversion via VSM method.

Table 2.2: The Implementation of VSM (Abual-Rub *et al*, 2007)

D_i = Denotes as document

Q_1 = Query

D_1: "shipment of gold damaged in a fire"

D_2: "Delivery of silver arrived in a silver truck"

D_3: "Shipment of gold arrived in a truck"

Q_1: "Gold silver truck"

$D=3$, $IDF=\log(D/df_i)$

Bag of Words	**Count tf_i**							$w_i = tf_i * idf_i$			
Terms	**Q**	**D_1**	**D_2**	**D_3**	**df_i**	**D/df_i**		**Q**	**D_1**	**D_2**	**D_3**
A	0	1	1	1	3	3/3=1		0	0	0	0
Arrived	0	0	1	1	2	3/2=1.5		0	0	0.1761	0.1761
Damage	0	1	0	0	1	3/1=3		0	0.4771	0	0
Delivery	0	0	1	0	1	3/1=3		0	0	0.4771	0
Fire	0	1	0	0	1	3/1=3		0	0.4771	0	0
Gold	1	1	0	1	2	3/2=1.5		0.1761	0.1761	0	0.1761
In	0	1	1	1	3	3/3=1		0	0	0	0
Of	0	1	1	1	3	3/3=1		0	0	0	0
Silver	1	0	2	0	1	3/1=3		0.4771	0		0
Shipment	0	1	0	1	2	3/2=1.5		0	0.1761	0	0.1761
Truck	1	0	1	1	2	3/2=1.5		0.1761	0	0.1761	0.1761

The VSM transforms every term inside the text documents into numerical values by defining them independently. Moreover, the terms are separated and being calculated based on their frequency appeared in text document. Treating all terms independently might cause misinterpretation on the content of the text document. This misinterpretation occurs when all terms are taken into account and being calculated the frequent occurred. This calculation is made without considering the term concepts such as Polysemy and Synonymy concepts which are identically significant. Nonetheless, VSM is often used in text document clustering in utilizing Natural Language Processing (NLP) where text documents are objects that represent human in expressing something via documentation.

In mathematical way VSM is represented with this Equation 2.1:

$$w_i = tf_i * \log\left(\frac{D}{df_i}\right) \quad (2.1)$$

Equation 2.1 shows the fundamental of weighting scheme when w_i denotes as total weight in numerical and tf_i is term frequency in a document. The method $\log\left(\frac{D}{df_i}\right)$ is used in order to normalize the value from $\frac{D}{df_i}$ which is D is documents and df_i is the document frequency that containing term i. With $\frac{D}{df_i}$ is the global probability due to its capability of choosing document contained terms related in other text documents.

Based on Table 2.2, it shows how the term frequencies are calculated. Assume that 3 documents (D_1, D_2, D_3) consisting terms is weighted based on $w_i = tf_i * idf_i$ and well converted into numerical values as per Table 2.2. All terms occurrences inside documents has been calculated their frequencies of occurred by using VSM formulation. All text documents will be generated vector values and these vector values determined how frequent of term appeared in text document based on the uppermost values. Figure 2.3 shows the example of BOW implementation, similar with VSM, which computes all terms occurrences in all text documents. With the samples of documents given in Documents D_1 and D_2 were extracted into several term occurrences with an elimination of repetitive terms into the list of terms. Nine entry vectors created from the distinct term collection. As a result D_1 and D_2 become a vector with numerical values of term occurrence and ignored the term relationship.

D_i = Document

D_1= {'Halim','likes','to','watch','the','football','games'}

D_2= {'Haizan','likes','to','watch','the','football','games','too'}

Extract single terms from D1 and D2 and construct a list of term:

Listed = {Halim:1, likes:2, to:3,watch:4, the :5, football:6, games:7, Haizan:8, too:9}

Creating 9 entry vectors based on distinct term collection:

D_1 = [1 1 1 1 1 1 1 0 0]

D_2 = [0 1 1 1 1 1 1 1 1]

Figure 2.3: The Implementation of BOW

2.4 The Vector Space Model (VSM) Limitation

Fundamentally, VSM treats the terms independently and may cause a major constraint by demolishing the relation between the terms, where terms can be comprehended as the conceptual representations, namely Polysemy and Synonymy. The first limitation, Polysemy, can be described as a term to express different things in different contexts, for example, "*collecting a card"* and "*collecting a result"*. Due to this, some unrelated documents might have high similarities because they tend to share some terms. As a result, this circumstance can affect the precision in the measurement issue. The second limitation, Synonymy, refer to a term used to express the same thing, such as "*truck driver"* and "*lorry driver"*. However, the number of similarities among some relevant documents might be low, as they do not share the same terms. This condition will effect on recall, in which these two limitations can be the detrimental causes to the vector created.

These limitations are exactly occurred when the process of converting text document is conducted by utilizing VSM formulation and along the process will discovered the limitations of VSM. From Table 2.2, the terms are separated and their relationship is ignored. This particular event shows a major limitation of VSM, where it works by making terms turned into form of single operation. Furthermore, these independencies will cause the ignorance of the relationship among terms which is important to create a possibility to represent the expression or meaning from text documents contents. For instance, the wrong interpretation of the contents can cause text document clustering less accurate while performing text clustering. The two major boundaries, Polysemy and Synonymy will be explained in further detail in the subtopics of Synonymy and Polysemy concepts respectively.

2.4.1 Synonymy Concept

Synonymy concept is very important to be highlighted in VSM issue, because it may cause the misinterpretation of text document content if it does not taken into account. This major issue happened while the process of text documents being converted into the sequences of numerical values. During the process, VSM ignore the term synonym concept and this ignorant will affected on the vector values. The effect to the vector values happened because of VSM threats the term inside document by

independently and without term relationship to be counted. Table 2.3 illustrates how Synonymy changes the text document content.

Table 2.3: Synonymy Effects in Text Documents

Text Document 1 (Synonymy Concept)			**Text Document 2 (VSM)**	
Terms	Frequency		Terms	Frequency
Sad, sorrow, Unhappy	{Sad, Sorrow, Unhappy} = 3		Sad, sorrow, Unhappy	Sad = 1, Sorrow = 1 and Unhappy = 1

The separation of terms in Text Document 2 in Table 2.3 has shown how synonym terms are treated by VSM as singleton and different to each other. This approach will cause the misinterpretation of text document content.

2.4.2 Polysemy Concept

According to Miller (1995), a word that has more than one sense is polysemous and two words that share at least one sense in common are synonymous such as phrases “driving a car” and “drives me crazy”. If both phrases are implemented in VSM, they will be separated into an individual term “driving”, “me”, “a”, “drives”, “car”, “ crazy” so the ignorance will take placed and change the meaning of text document. As a result, VSM affects the content of text document by separating terms into singleton. At the end of the spectrum, this limitation can be countered by detecting the phrases in text documents in order to look after the term contact. The term contact is really significant because the term contact may have the Polysemy concept that lies inside the text documents that truly believed will give some improvement to the text document clustering. Table 2.4 below shows how this Polysemy concept is significant to figure the text document content.

Table 2.4: Polysemy Effects in Text Documents

Text Document 1 (Polysemy Concept)			Text Document 2 (VSM)	
Terms	Frequency		Terms	Frequency
My name is Gabriel	{My name is Gabriel} = 1		My name is Gabriel	My = 1, Name = 1 is = 1 Gabriel = 1

The Polysemy concept will keep term contact in default position and it protects the originality of the meaning. Such as in Table 2.4 the Polysemy concept is still a phrase of "My name is Gabriel".

2.5 Text Document Clustering Based on Frequent Concept (WordNet)

Terms are important in representing passage and passage is important to represents document. In this research term frequent detection method is used in order to investigate the document relationship existed inside text documents. In order to consider the term concept, it is necessary to have an electronic lexical dictionary as term concept database. Since WordNet is choose in this study, a further research on WordNet dictionary is conducted.

WordNet has been used widely in many text documents' clustering research due to its capacity of Synsets word groups. Furthermore, WordNet is also well known as a large lexical database with the combination of a dictionary and a thesaurus of English language (Miller, 1995). Furthermore, Miller (1995), defines the vocabulary of a language as a set W of pairs (f, s), where a form f is a string over a finite alphabet, and a sense s is an element from a given set of meanings. For instance, forms can either be utterances composed of a string of phonemes or inscriptions composed of a string of characters. Due to this, each form with a sense in a language is called a term in that language. Other than that, a dictionary is referred as an alphabetical list of words, in which a word that has more than one sense is polysemous while two terms that share at least one sense in common are said to be synonymous (Miller, 1995).

In addition, the purposeful sentences usually consist of meaningful terms. Any approach in computing Natural Language Processing (NLP) must have detailed features of information about terms and their meanings. In order to obtain the details,

an ordinary dictionary is normally the solution but it is not machine interpreted or machine readable. A dictionary is suitable for human only and to initiate machine learning requires a specific version in computer readable. However, instead of the ordinary dictionary, WordNet was invented. WordNet is an online lexical database designed for the use under program control. English nouns, verbs, adjectives, and adverbs are organized into sets of synonyms, in which each represents a lexicalized concept and a semantic relation link the synonym sets. In conjunction with this research, both concepts of Polysemy and Synonymy are the main keys to investigate how both concepts are able to mend the limitations of VSM as mentioned previously by defining the similarity between text documents with the help of Synsets. Synsets is a group of synonymous words and collocations and it corresponds to a concept (Baghel & Dhir, 2010). Synsets came along with WordNet which is contains 155 287 words organized in 117 659 Synsets for a total of 206 941 word sense with pair (Baghel & Dhir, 2010).

In WordNet, there are semantic relations among terms and these relations are divided into six categories (Miller, 1995). As shown in Figure 2.4 is a concept of relationships that consists of useful tools such as Synonymy, Antonymy, Hyponymy, Meronymy, Troponymy and Entailment. However, this research only focuses on Synonymy concept to counter VSM limitation.

Semantic Relation	Syntactic Category	Examples
Synonymy (similar)	N, V, Aj, Av	pipe, tube rise, ascend sad, unhappy rapidly, speedily
Antonymy (opposite)	Aj, Av, (N, V)	wet, dry powerful, powerless friendly, unfriendly rapidly, slowly
Hyponymy (subordinate)	N	sugar maple, maple maple, tree tree, plant
Meronymy (part)	N	brim, hat gin, martini ship, fleet
Troponomy (manner)	V	march, walk whisper, speak
Entailment	V	drive, ride divorce, marry

Note: N = Nouns Aj = Adjectives V = Verbs Av = Adverbs

Figure 2.4: Semantic Relations in WordNet

Many researchers (Huang *et al.,* 2008; Hamou *et al.*, 2010; Ray & Singh, 2010; Thanh & Yamada, 2011; Bouras & Tsogkas, 2012; Celik & Gungor, 2013) have used WordNet as a platform to extract word ontology and the text extraction feature to apply on their datasets in finding Synonymy concept. Baghel & Dhir (2010), has proposed Frequent Concept Document Clustering (FCDC), which is contained WordNet as references for concept of Synonymy to cluster text document. In this approach, Baghel & Dhir (2010) has stated the solution by searching for the concepts in documents and then finding for the frequent concepts. Baghel & Dhir (2010) applied the concepts by using apriori paradigm that utilizes frequent item set. Next, the paradigm forms the initial clusters of documents with each cluster representing a single frequent concept. Then, the proposed algorithm processes these initial clusters to create disjoint clusters and finally represents the results by using the hierarchical tree like structure.

Furthermore, Abdelmalek, Elberrichi & Simonet (2010) have deployed WordNet as a research platform in order to extract the term concept before applying clustering algorithm. In experimental process, WordNet was used as text document representation and the representation with WordNet is better because it make the similarity among text document increased. Moreover, the text clustering used an ascending hierarchical clustering method known as a Self-Organizing Maps (SOM) based clustering method and an ant-based clustering method. These algorithms were applied based on the Synsets of WordNet as terms for the representation of textual documents. As a result, the effects of these methods were examined in several experiments using three similar measurements, namely the Cosine distance, the Euclidean distance and the Manhattan distance. Other than that, the experiments indicated that using the SOM based clustering method and using the cosine distance provided the best results. Abdelmalek *et al.* (2010) has mentioned in a publication that many document clustering tasks use Synsets of WordNet as a featured selection in order to gain accuracy of the clustering results.

Bouras & Tsogkas (2012) have investigated the usefulness of WordNet to improved text document clustering quality (accuracy). The accuracy had achieved by the utilization of the hypernyms concept obtained from WordNet database. This hypernyms is a linguistic term for a word whose meaning includes the meaning of other words. As example, fish is hypernyms of tuna and dolphin. In another words, hypernyms (also called superordinates) are general words hyponyms are subdivisions

of more general words. The term concept (hypernyms) was used in the preprocessing step and applied on several text clustering algorithms.

The research was proposed W-K means (K-means WordNet based) algorithm and compared it with K-means. The comparison between both clustering algorithms have resulted the proposed one better than K-means, where it was improved by the enhancement of standard K-means algorithm using the external knowledge from WordNet. Moreover, the experimental process used the approach of WordNet hypernyms in a two folded manner, enriching the ‘‘bag of words’’ prior to the clustering process and assisting the label generation procedure. In addition, both processes were implemented in text preprocessing phase. Thus, the research shown how WordNet becomes very useful to increase or improve the clustering algorithms.

On the other hand, the additional WordNet dictionary before the clustering process also can be used as a feature selection. It is conjunction with huge increased in the number of text documents on the Internet nowadays and the fast speed material publication such as news article, digital libraries, publications and many more, this consequence has yielded a major problem of high dimensionality of data that which is very high volume of data.

Patil & Atique (2013) have investigated how feature selection improves the text classification in order to deal with the major problem as mentioned. The research were started with the used of feature selection method from the preprocessing steps on text documents by selecting the important feature with the information gained based on entropy. Based on WordNet as a lexical database, the global unique word has been extracted and found the frequent word sets and further key terms from all documents by using feature selection methods. From the results, the important terms were selected based on the information gained from using WordNet was reduced the text document dimensionality.

The useful of WordNet is well proven since it was used by many researchers (Huang *et al.,* 2008; Hamou *et al.,* 2010; Ray & Singh, 2010; Thanh & Yamada, 2011; Bouras & Tsogkas, 2012; Celik & Gungor, 2013). In addition, this lexical dictionary provided Synsets (synonym sets) used for combining single terms into term concept to become a more significant form. This advantage is very useful to overcome VSM’s limitation by utilising WordNet dictionary in order to concatenate the terms that related to the Synonymy concept.

2.6 Text Document Clustering Based on N-grams

N-grams is another method of text document representation besides Vector Space Model (VSM) which has several advantages ("N-grams" is a sequence of n consecutive characters or term) (Amine *et al*., 2008). The whole set of N-grams (n generally varies from 2 to 5) generated for a given document is mainly the result of the displacement of a window of *n* equal to characters along the text (Millar *et al*., 2006). Furthermore, the window is moved by a character at a time and the position. This method creates term N-grams of tokens in a document as based on Markov Model (Brown *et al*., 1992). Other than that, *n* can be assigned in view of term or character (linguistic area of study), where *n* = term or *n* = *character*. These connotations define term N-grams as a series of consecutive tokens of length *n*. The term N-grams is generated by consisting of all series of consecutive tokens of length *n*. In an experiment by Shannon *et al*. (1949) *n* was located with probability when language modeling and independence assumptions are made so that each word depends only on the last *n*-1 words. This N-grams method can be used in many areas of study widely such as Protein Sequencing, DNA Sequencing, Computational Linguistics (character) and Computational Linguistics (term). For instance, this study only focuses on the use of Computational Linguistic (term) and works with text document clustering. Table 2.5 shows in general how N-grams method converts sequence of term into consecutive relative term, merging the same properties together. With this N-grams ability, it helps to solve VSM problems on the Polysemy limitation by concatenating terms inside text documents.

Table 2.5: Generated N-grams Term

Domain	Unit	Sample	1-gram	2-gram	3-gram
Computational linguistics	term	me or me not to me	me, or, me, not, to, me,	me or, or me, me not, not to, to me,	me or me, or me not, me not to, not to me,

Cavnar (1995) used N-grams based representations for text documents. N-grams is deployed by Cavnar (1995) because of several distinct advantages for various document processing tasks. First, N-grams provide a more robust representation in

the face of grammatical and typographical errors in the documents. Secondly, N-grams representations require no linguistic preparations such as word-stemming or stopwords removal. Thus the N-grams are ideal in situation requiring multi language operations. Vector processing retrieval models also have some unique advantages for information retrieval tasks. In particular, based on N-grams, Cavnar (1995) has provided a simple and uniform representation for documents queries, and an intuitively appealing document similarity measure. Besides that, the technique of modern vector space model has good retrieval performance by characteristics. In addition Cavnar (1995) worked on the combination of these two ideas by using a vector processing model for documents and queries based on N-grams frequencies as the basis for the vector element values instead of more traditional term frequencies. The resulting system provides good retrieval performance on the TREC-1 and TREC-2 tests dataset without the need for any kind of word stemming or stopwords removal. Furthermore Cavnar (1995) also had begun testing the system on other language documents (Spanish language) rather than English.

Miao, keselj & Milios (2005) had motivated by some recent positive results in using character N-grams in building author profiles and used it in automated authorship attribution. Furthermore Miao *et al.* (2005) explored the idea of using frequent character N-grams as vector features in document clustering. Based on N-grams, a vector space has been built for the documents and found that the results are better than using terms, especially for low dimensionality. Miao *et al.* (2005) also had done an experiment with multi word terms as the vectors features. Based on the hypothesis is on the assumption that a multi-word term representation is a more compact representation of meaning in a domain than words, and therefore has the potential of providing better clustering performance at lower dimensionality. They perform automatic term extraction based on a combination of linguistic and statistical criteria. The clustering algorithm been used in the experimental process is K-Means with cosine distance measure.

The approach of N-grams had been proposed by Kim *et al*. (2010) as dynamic modelling information within the text data like audio modelling scenario for information retrieval applications. Purposely, Kim *et al*. (2010) used the bigram model to consider adjacent acoustic words and built a new acoustic word dictionary for the bigrams. Experimental results showed that the proposed N-grams approach brought a significant improvement in the performance by providing complementary

REFERENCES

Abdelmalek Amine., Zakaria Elberrichi., Michel Simonet, (2010) Evaluation of Text Clustering Methods Using WordNet, The International Arab Journal of Information Technology, Vol. 7, No. 4, pp. 349-355.

Abual-Rub, M. S., Abdullah, R., & Rashid, A. (2007). A Modified Vector Space Model for Protein Retrieval. International Journal of Computer Science and Network Security (IJCSNS), 7(9), 85-89.

Alneyadi, S., Sithirasenan, E., &Muthukkumarasamy, V. (2013). Word N-grams Based Classification for Data Leakage Prevention. In Trust, Security and Privacy in Computing and Communications (TrustCom), 2013 12th IEEE International Conference, pp. 578-585.

Amine, A., Elberrichi, Z., Simonet, M., & Malki, M. (2008). Wordnet-based and N-grams-based document clustering: A comparative study. InBroadband Communications, Information Technology & Biomedical Applications, 2008 Third International Conference, pp. 394-401.

Baghel, R., & Dhir, R. (2010). Text document clustering based on frequent concepts. In Parallel Distributed and Grid Computing (PDGC), 2010 1st International Conference, pp. 366-371.

Beil, F., Ester, M., & Xu, X. (2002). Frequent term-based text clustering. InProceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 436-442.

Bengio, Y., Ducharmea R., Vincent P., & Jauvin C. (2006). Neural probabilistic language models. Innovations in Machine Learning, pp. 137-186.

Bharati & M. Ramageri (2010). Data Mining Techniques and Applications. Indian Journal of Computer sciences and Engineering, Vol.1 No.4, pp. 301-305.

Bouras, C., & Tsogkas, V. (2012). A clustering technique for news articles using WordNet. Knowledge-Based Systems, pp. 115-128.

Brown, Desouza, Mercer, Della Pietra, Jenifer C Lai (1992). Class-based n-gram models of natural language.Computational linguistics, pp.467-479.

Buscaldi, D., Tournier, R., Aussenac-Gilles, N., & Mothe, J. (2012). Irit: Textual similarity combining conceptual similarity with an N-grams comparison method. In Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation Association for Computational Linguistics, pp. 552-556.

Callison Burch, Chris, and Raymond S. Flournoy. (2001) A program for automatically selecting the best output from multiple machine translation engines. Proc. of MT Summit VIII. pp 3-5.

Celik, K., & Gungor, T. (2013). A comprehensive analysis of using semantic information in text categorization. In Innovations in Intelligent Systems and Applications (INISTA), 2013 IEEE International Symposium, pp. 1-5.

Carlone, D. (1989). *Sentiment Mining in a Location-Based Social Networking Space.* University of Aalborg: Master. Thesis.

Cavnar, W. (1995). Using an N-grams-based document representation with a vector processing retrieval model. NIST SPECIAL PUBLICATION SP, pp. 269-269.

E. Statmatatos (2011). Plagiarism Detection Using Stopword n-grams. Journal of American Society for Information Science and. International Journal of Computer Science Issues, Vol. 62, Issue 12, pp 2512-2527.

Elahi, A., & Rostami, A. S. (2012). Concept-based vector space model for improving text clustering. Journal of Advanced Computer Science & Technology Research. pp 2-3.

Elberrichi, Z., Rahmoun, A., & Bentaallah, M. A. (2008). Using WordNet for Text Categorization. Int. Arab J. Inf. Technol., 5(1), pp. 16-24.

Gergely Pethö. (2001) What is Polysemy A Survey of Current Research and Results. In: Enikő Németh T., Károly Bibok: Pragmatics and the Flexibility of Word Meaning. Elsevier, Amsterdam, pp. 175–224.

Go, K., & See, S. (2008). Incorporation of WordNet Features to N-grams Features in a Language Modeler. In PACLIC, pp. 179-188.

Hamou, R. M., Lehireche, A., Lokbani, A. C., & Rahmani, M. (2010). Representation of textual documents by the approach wordnet and n-grams for the unsupervised classification (clustering) with 2D cellular automata: a comparative study. Computer and Information Science, 3(3), pp. 240-255.

Huang, A. (2008). Similarity measures for text document clustering.In Proceedings of the sixth New Zealand computer science research student conference (NZCSRSC2008), Christchurch, New Zealand, pp. 49-56.

Huang, A., Milne, D., Frank, E., & Witten, I. H. (2009). Clustering documents using a Wikipedia-based concept representation. In Advances in Knowledge Discovery and Data Mining Springer Berlin Heidelberg, pp. 628-636.

Kalaivendhan, K., & Sumathi. (2014) M. P. International Journal of Innovative Research in Engineering Science and Technology http://journal. selvamtech.com. pp 257-266.

Kim, S., Sundaram, S., Georgiou, P., & Narayanan, S. (2010). An N-grams model for unstructured audio signals toward information retrieval. InMultimedia Signal Processing (MMSP), 2010 IEEE International Workshop,pp. 477-480.

Liu, C., Wang, D., & Tejedor, J. (2012). N-grams FST Indexing for Spoken Term Detection. In interspeech.

MacQueen, James. (1967.) Some methods for classification and analysis of multivariate observations." Proceedings of the fifth Berkeley symposium on mathematical statistics and probability. Vol. 1, pp. 281-297.

Miao, Y., Kešelj, V., & Milios, E. (2005). Document clustering using character N-grams: a comparative evaluation with term-based and word-based clustering. In Proceedings of the 14th ACM international conference on Information and knowledge management ACM, pp. 357-358.

Millar, E., Shen, D., Liu, J., & Nicholas, C. (2006). Performance and scalability of a large- scale n-gram based information retrieval system. Journal of digital information, 1(5).

Miller, G. A. (1995). WordNet: a lexical database for English. Communications of the ACM, 38(11), pp. 39-41.

N. C. Thanh. & K. Yamada (2011). Document Representation and Clustering with WordNet Based Similarity Rough Set Model." International Journal of Computer Science Issues, Vol. 8, Issue 5, No 3.

Patil, L., & Atique, M. (2013). A novel feature selection based on information gain using WordNet. In Science and Information Conference (SAI), pp. 625-629).

Ravichandra Rao, I. K. (2003). Data mining and clustering techniques.DRTC Workshop on Semantic Web 8th – 10th December DRTC, Bangalore. Pp. 325-330.

Ray, S. K., & Singh, S. (2010). Blog content based recommendation framework using WordNet and multiple ontologies. In Computer Information Systems and Industrial Management Applications (CISIM), 2010 International Conference, pp. 432-437.

Salton, G., Wong, A., & Yang, C. S. (1975). A vector space model for automatic indexing. Communications of the ACM, 18(11), pp. 613-620.

Shannon, Claude Elwood, Warren Weaver, and Richard E. Blahut (1949). The mathematical theory of communication. Vol. 117. Urbana: University of Illinois press.

Srividhya, V., & Anitha, R. (2011). Evaluating preprocessing techniques in text categorization. International Journal of Computer Science and Application,47(11).

Steinbach, M., Karypis, G., & Kumar, V. (2000). A comparison of document clustering techniques. In KDD workshop on text mining Vol. 400, No. 1, pp. 525-526.

Suyal, H., Panwar, A., & Singh Negi, A. (2014). Text Clustering Algorithms: A Review. International Journal of Computer Applications, 96(24), pp. 36-40.

Toby S,Collin E and jammy T (2009) in :Data Integration Across the Web. Programmming the Semantic Web. O'Reilly Media, Inc.1005 Gravenstein Highway North Sebastopol, CA 95472 800-998-9938 (in the United States or Canada).

Vaquero, A., Sáenz, F. and Barco, A., (2000), Computer-based tools for improving the language mastery: Authoring and using electronic dictionaries, V Congreso Iberoamericano de Informática Educativa, RIBIE 2000, Vina del Mar, Chile.

Yu Xiao, (2010) A Survey of Document Clustering Techniques & Comparison of LDA and moVMF Comparison of LDA and moVMF ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1-4.

Zhao, Y., Karypis, G., & Fayyad, U. (2005). Hierarchical clustering algorithms for document datasets. Data Mining and Knowledge Discovery, 10(2), 141-168.

ZhiLiu (2011). Uci Machine Learning Repository: Retrieved from http://archive.ics.uci.edu/ml.

Zhou, A. W., & Yu, Y. F. (2011). The Research about Clustering Algorithm of K-Means. Computer Technology and Development, 21(2), pp. 62-65.