ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 1, Issue 7, September 2013

An Efficient Approach in Text Clustering Based on Frequent Itemsets

S.Murali Krishna, S.Durga Bhavani

Associate Professor, S.V. Collge of Engineering, A.P, India

Professor, Department of Computer Science and Engineering School of Information, Technology (SIT), JNTU-Hyderabad, A.P, India

ABSTRACT: In recent times, the vast amount of textual information available in electronic form is growing at staggering rate. This increasing number of textual data has led to the task of mining useful or interesting frequent itemsets (words/terms) from very large text databases and still it seems to be quite challenging. The use of such frequent itemsets for text clustering has received a great deal of attention in research community since the mined frequent itemsets reduce the dimensionality of the documents drastically. In the proposed research, we have devised an efficient approach for text clustering based on the frequent itemsets. A renowned method, called Apriori algorithm is used for mining the frequent itemsets. The mined frequent itemsets are then used for obtaining the partition, where the documents are initially clustered without overlapping. Furthermore, the resultant clusters are effectively obtained by grouping the documents within the partition by means of derived keywords. Finally, for experimentation, the Reuter-21578 dataset are used and thus the obtained outputs have ensured that the performance of the proposed approach has been improved effectively.

KEYWORDS: Text mining, Text clustering, Text documents, Frequent itemsets, Apriori, Reuter-21578.

1. INTRODUCTION

The rapid progress of databases in every aspect of human actions has resulted in enormous demand for efficient tools for turning data into valuable knowledge. In order to fulfill this requirement, researchers from numerous technological areas, namely pattern recognition, machine learning, data visualization, statistical data analysis, neural networks, information retrieval, econometrics, information extraction etc., have been searching for eminent approaches. The entire efforts have resulted in an effective research area known as data mining (DM) or Knowledge Discovery in Databases (KDD) [6]. Commonly, data mining is carried out on data represented in quantitative, multimedia or textual forms [21]. But, at present, a large amount of information exists in the model of text, comprising of documents, news, email, manuals and more. The access to a large quantity of textual documents turns out to be effectual because of the growth of the digital libraries, web, technical documentation, medical data and more. These textual data comprise of resources which can be utilized in a better way. Thus,knowledge discovery from textual databases, otherwise termed as, text mining (TM), is a prominent and tough challenge, due to the value and ambiguity of natural language which is employed in majority of the existing documents [1]. Text mining is a major research field due to the need of acquiring knowledge from the large number of available text documents, particularly on the Web [8].

Both text mining and data mining are part of information mining and identical in some perspective. Data mining techniques can be adapted in a better way to mine text [8]. In the two knowledge-discovery variants, when the data is organized, document elements are tagged or numerical data is represented in organized data structures. There exists an identical application of statistical techniques to lessen the complexity of the problem, to recognize correlations, linkages, clusters, and relationships, and to make predictive rules, which are also identified in few circles as knowledge [11, 12, and 13]. Text mining can be described as a knowledge-intensive process in which a user communicates with a collection of documents in due course by employing a set of analysis tools. Similar to data mining, text mining anticipates mining valuable information from data sources by recognition and searching of interesting patterns. “Text mining” refers to the application of data mining techniques to automated discovery of valuable or interesting

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 1, Issue 7, September 2013

information from unstructured text [4], [3], [2] and [10].

Text mining is a progressively more significant research field since the requirement of attaining knowledge from the massive amount of text documents [34]. In order to mine large document collections, it is vital to pre-process the text documents and save the data in a data structure, which is suitable for processing it further than a plain text file [18]. Typically, text preprocessing includes tokenization, Part of Speech (PoS) Tagging [19], word stemming and the application of a stop words removal technique. Tokenization is referred as the procedure of splitting the text into words or terms. Using Part of Speech Tagging, words can be tagged according to the grammatical context of a word in the sentence and there by words can be divided into nouns, verbs and more [20]. Information Extraction is defined as the mapping of natural language texts (namely newswire reports, journal and newspaper articles, World Wide Web pages, electronic mail, any textual database and more.) into predefined structured representation, or templates, which, when filled, represent an extract of key information from the original text [7]. Of late, extracting relationships from entities in text documents has achieved significant interest. Association rule mining discovers the interesting associations and/or correlation relationships among large set of data items.

Mining association rules in transaction databases has been demonstrated to be valuable in a wide range of application areas [15, 16]. Yet, its application on text databases still seems to be more promising, owing to the difference in characteristics of transaction databases with text databases [14]. In the case of text mining, extracted rules are deduced as co-occurrences of terms in texts and therefore are able to return semantic relations among the terms [17]. Text mining is a multidisciplinary field, which includes these functions: text analysis, information retrieval, clustering, information extraction, categorization, visualization, machine learning, data mining and database technology [22]. The method of dividing data objects (e.g.: document and records) into significant clusters or groups such that objects within a cluster possess analogous characteristics but are contradictory to objects in other clusters is known as Cluster analysis [4], [5]. Text clustering is defined as an efficient way for sorting several documents to assist users navigate, summarize, and arrange text documents [31, 32, 9]. By arranging a huge number of documents into meaningful clusters, document clustering can be employed to browse a set of documents or to arrange the results given by a search engine in answer to a user's query. This can considerably enhance the accuracy and recall in information retrieval systems, and it is a proficient way to determine the nearest neighbors of a document [33].

In this paper, we have presented an effective frequent itemset-based document clustering approach. First, the text documents in the text data are preprocessed with the aid of stop words removal technique and stemming algorithm. Then, top- *p* frequent words are extracted from each document and

hence, we form the binary mapped database through the use of extracted words. We apply the Apriori algorithm to discover the frequent itemsets having different length. The mined frequent itemsets are sorted in descending order based on their support level for every length of itemsets. Subsequently, we split the documents into partition using the sorted frequent itemsets. These frequent itemsets can be viewed as understandable description of the obtained partitions. Furthermore, the resultant cluster is formed within the partition using the derived keywords (words obtained by taking the absolute complement of familiar keywords with respect to the top- *p* frequent words).

The basic outline of this paper is as follows. Section 2 describes the brief review of related researches of text clustering. Section 3 presents the proposed approach for text clustering. Section 4 describes the experimental results and the performance evaluation of the proposed approach. Section 5 gives conclusion of the proposed approach.

II. REVIEW OF RELATED RESEARCHES

A handful of researches are available in the literature for clustering the text data in which, frequent itemsets based text clustering have received a lot of attention among the researchers recently. A brief

review of some recent researches related to frequent itemsets-based text clustering is presented here.

Zhou Chong *et al.* [23] have presented a method known as Frequent Itemset-based Clustering with Window (FICW), which employed the semantic information for text clustering with a window constraint. The experimental results obtained from three (hypertext) text sets revealed that FICW performed better in terms of both clustering accuracy and efficiency. Xiangwei Liu and Pilian He [24] have introduced a text-clustering algorithm known as Frequent Term Set-based Clustering (FTSC). It employs frequent term sets to cluster texts. Initially, it extracts significant information from documents and put it into databases. Later, it employed the Apriori to mine the frequent item sets. At last, it clusters the documents as per the frequent words in subsets of the frequent term sets. The algorithm can lessen the dimension of the text data for extremely large databases, so it could enhance the accuracy and speed of the clustering algorithm. The experimental results showed that FTSC and FTSHC algorithms are

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 1, Issue 7, September 2013

comparatively more efficient than K-Means algorithm in the clustering performance.

Le Wang *et al.* [25] have presented a simple hybrid algorithm (SHDC) on the basis of top-k frequent term sets and k-means so as to overcome the main challenges of current web document clustering. Top-k frequent term sets were employed to provide k initial means, which were regarded as initial clusters and later refined by k-means. The final optimal clustering was returned by k-means whereas the clear description of clustering was given by *k* frequent term sets. Experimental results on two public datasets showed that SHDC performed better other two representative clustering algorithms (the farthest first k-means and random initial k-means) both on efficiency and effectiveness. Zhitong Su *et al.* [26] have introduced a web-text clustering method for personalized e-learning based on maximal frequent itemsets. In the beginning, the Web documents were represented by vector space model. Later, maximal frequent word sets were determined. In the end, on the basis of a new similarity measure of itemsets, maximal itemsets were employed for clustering documents. Experimental results proved that the presented method was efficient.

Yongheng Wang *et al.* [27] have introduced a frequent term based parallel clustering algorithm which could be employed to cluster short documents in very large text database. A semantic classification method is also employed to enhance the accuracy of clustering. The experimental analysis proved that the algorithm was more precise and efficient than other clustering algorithms when clustering large scale short documents. In addition, the algorithm has good scalability and also could be employed to process huge data. The documents clustering algorithm on the basis of frequent term sets was proposed by W.L. Liu and X. S. Zheng [29]. Initially, documents were denoted as per the Vector Space Model (VSM) and every term is sorted in accordance with their relative frequency. Then frequent term sets can be mined using frequent-pattern growth (FP growth). Lastly, documents were clustered on the basis of these frequent term sets. The approach was efficient for very large databases, and gave a clear explanation of the determined clusters by their frequent term sets. The efficiency and suitability of the proposed algorithm has been demonstrated with the aid of experimental results.

Henry Anaya-Sanchez *et al.* [30] have proposed a clustering algorithm for discovering and unfolding the topics included in a text collection. The algorithm depended on the most probable term pairs generated from the collection and also on the estimation of the topic homogeneity related to these pairs. Topics and their descriptions were produced from those term pairs whose support sets were homogeneous for denoting collection topics. The obtained experimental results over three benchmark text collections showed the efficacy and usefulness of the approach. Florian Beil *et al.* [28] have proposed an approach which employed frequent item (term) sets for text clustering. Such frequent sets were determined by means of algorithms for association rule mining. So to cluster on the basis of frequent term sets, they calibrated the mutual overlap of frequent sets with regard to the sets of supporting documents. They provided two algorithms for frequent term-based text clustering, FTC which produced flat clustering and HFTC for hierarchical clustering. An experimental assessment on classical text documents and also on web documents showed that the presented algorithms obtain clustering of comparable quality appreciably more efficiently than modern text clustering algorithms.

III. AN EFFICIENT APPROACH FOR TEXT CLUSTERING BASED ON FREQUENT ITEMSETS

The reputation of the Web and the huge quantity of documents existing in electronic form has provoked the exploration for hidden knowledge in text collections. Therefore, there is an increasing research concentration in the general topic of text mining. For finding the meaningful information from the text documents, researchers have used various data mining techniques, in which clustering is one of the popular technique. Text clustering is to group a collection of documents (unstructured texts) into different category groups so that documents in the same category group describe the same subject. Many researches [23-26, 28] have investigated possible ways to improve the performance of text document clustering based on the popular clustering algorithms (partitional and hierarchical clustering) and frequent term based clustering. Here, we have devised an effective approach for clustering a text corpus with the aid of frequent itemsets. The devised approach consists of the following major steps:

1) Text preprocessing
2) Mining of frequent itemsets
3) Partitioning the text documents based on frequent Itemsets
4) Clustering of text documents within the partition

A.TEXT PREPROCESSING

Let D be a set of text documents represented as $D = \{d_1\ d_2\ d_3 .....d_n\}; 1 \le i \le n$, where, n is the number

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 1, Issue 7, September 2013

documents in the text dataset D. The text document set D is converted from unstructured format into some common representation using the text preprocessing techniques, in which the words or terms are extracted (tokenization). The input data set D (text documents) are preprocessed using the techniques namely, removing stop words and stemming algorithm.

a) *Stop word Removal:* Removes the stop (linking) words like “have”, “then”, "it", "can”, “need", "but", "they", "from", "was", "the", "to", "also" from the document [36].
b) *Stemming algorithm:* Removes the prefixes and suffixes of each word [35].

B. MINING OF FREQUENT ITEMSETS

This sub-section describes the mining of frequent itemsets from the preprocessed text documents D. For every document d_i, the frequency of the extracted words or terms from the preprocessing step is

computed and the top- p frequent words from each document d_i are taken out.

$$K_w = \{ d_i \mid p(d_i) \quad ; \quad \forall d_i \subseteq D \}$$

where, $p(d_i) = T_{wj} \quad ; \quad 1 \leq j \leq p$

From the set of top- p frequent words, the binary database B is formed by obtaining the unique words. Let B_T be a binary database consisting of n number of transactions (documents) T and q

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 1, Issue 7, September 2013

number of attributes (unique words) $U = [u_1, u_2, ..., u_q]$. Binary database B_T consists of binary data that represents whether the unique words are presented or not in the documents d_i.

$$B_T = \begin{cases} 0 & if \quad u_j \notin d_i \\ 1 & if \quad u_j \in d_i \end{cases} \quad ; \quad 1 \le j \le q, \quad 1 \le i \le n$$

Then, the binary database B_T is given to the Apriori algorithm for mining the frequent itemsets (words/terms) F_S.

3.2.1. Apriori Algorithm

Apriori is a conventional algorithm that was first introduced in [37] for mining association rules. The two steps used for mining association rules are as follows. (1) Identifying frequent itemsets (2) Generating association rules from the frequent itemsets. Frequent itemsets can be mined in two steps. At first, candidate itemsets are generated and afterwards frequent itemsets are mined with the help of these candidate itemsets. Frequent itemsets are nothing but the itemsets whose support is greater than the minimum support specified by the user. In the proposed approach, we have used only the frequent itemsets for further processing so that, we undergone only the first step (generation of frequent itemsets) of the Apriori algorithm. The pseudo code corresponding to the Apriori algorithm [38] is,

Pseudo code:

C_k : Candidate itemset of size k

I_k : Frequent itemset of size k.

```
I1 = {large 1-itemsets};
for (k = 2; Ik-1 ≠ 0; k++) do begin
        Ck = apriori-gen(Ik-1);          // New candidates
        for all transactions T ∈ D do begin
            CT = subset(Ck, T);  // Candidates contained in T
            for all candidates c ∈ CT do
                c.count++;
            end
        end
        Ik = {c ∈ Ck | c.count ≥ minsup}
end
```

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 1, Issue 7, September 2013

t Documents Based on Frequent Itemsets

This section describes the partitioning of text documents D based on the mined frequent itemsets F .

Definition1: *Frequent itemset* is a set of words that occur together in some minimum fraction of documents in a cluster. The Apriori algorithm generates a set of frequent itemsets with varying length (l) from 1 to k . First, the set of frequent itemsets of each length (l) are sorted in descending order in accordance with their support level.

$F_s = \{ f_1 \quad f_2$

$f_3 \ldots f_k \}$;

$1 \leq l \leq k$

$f_l = \{$

Answe

$r = \mathrm{U}_k$

I_k ;

3. 3. Partitioning the Tex

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 1, Issue 7, September 2013

$f_{l(i)}$; $1 \le i \le t\}$

where, set f_l . $^{\sup(f_{l(1)}) \ge \sup(f_{l(2)}) \ge \ldots \ge \sup(f_{l(t)})}$ and t denotes the number of frequent itemsets in the

Initially, the first element ($f_{(k/2)}(1)$) from the sorted list $f_{(k/2)}$, which is a set of frequent itemsets is selected. Subsequently, an initial partition c_1 , which contains all the documents having the itemset $f_{(k/2)}(1)$, is constructed. Then, we take the second element $f_{(k/2)}(2)$, whose support is less than $f_{(k/2)}(1)$ to form a new partition c_2 . This new partition c_2 is formed by identifying all the documents having frequent itemset $f_{(k/2)}(2)$ and takes away the documents that are in the initial partition c_1 . This procedure is repeated until every text documents in the input dataset D are moved into partition $C_{(i)}$. Furthermore, if the above procedure is not terminated with the sorted list $f_{(k/2)}$, then the subsequent sorted lists ($f_{((k/2)-1)}$, $f_{((k/2)-2)}$ etc..) are taken for performing the above discussed step (inserting the documents into the partition). This results a set of partition c and each partition $C_{(i)}$ contains a collection documents $D^{(x)}_{c(i)}$

$c = \{ c_{(i)} \mid c_{(i)} \in f_{l(i)} \}$ $1 \le i \le m$, $1 \le l \le k$

;

$C_{(i)} = Doc[f_{l(i)}]$; $C_{(i)} = \{ D^{(x)}_{c(i)};\ D^{(x)}_{c(i)} \in$ D , $1 \le x \le r$ $\}$

Where, m denotes the number of partitions and r denotes the number of documents in each partition.

For constructing initial partition (or cluster), we make use of mined frequent itemset which significantly reduces the dimensionality of the text document set and clustering with reduced dimensionality is considerably more efficient and scalable. The clustering results produced by the approaches presented in [41, 28] consist of the overlapping of documents due to the use of frequent itemsets and these overlapping documents have been removed to obtain the final results. In the proposed research, we directly generate the non-overlapping partitions from the frequent itemsets. This makes the initial partitions disjoint, because the proposed approach keeps the document only within the best initial partition.

3.4. Clustering of Text Documents within the Partition

In this sub-section, we describe how to cluster the set of partitions obtained from the previous step. This step is necessary to form a sub cluster (describing sub-topic) of the partition (describing same topic) and the resulting cluster can detect the outlier documents significantly. Furthermore, the proposed approach does not require a pre-specified number of clusters. The devised procedure for clustering the text documents available in the set of partition c is discussed below.

In this phase, we first identify the documents $(x$ and the familiar words $f_{c(i)}$ (frequent $D^{)}_{c(i}$ $)$

itemset used for constructing the partition) of each partition $C_{(i)}$. Then, the derived keywords K_d $[D^{(x)}_{c(i)}]$ of document $(x$ $)$ $D^{)}_{c(i}$

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 1, Issue 7, September 2013

are obtained by taking the absolute complement of family words $f_{c(i)}$ with respect to the top- p frequent words of the document $D_{c(i)}^{(x)}$.

$$K_d[D_{c(i)}^{(x)}] = \{T_{w_j \setminus f_{c(i)}} ; T_j \in D_{c(i)}^{(x)}, 1 \le i \le m, 1 \le j \le p, 1 \le x \le r\}$$

$$T_{w_j \setminus f_{c(i)}} = \{x \in T \mid x \notin f_{c(i)}\} \; w_j$$

The set of unique derived keywords of each partition $C_{(i)}$ are obtained and the support of each unique derived keyword is computed within the partition. The set of keywords satisfying the cluster support (cl_sup) are formed as representative words of the partition $C_{(i)}$. ***Definition2:*** The *cluster*

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 1, Issue 7, September 2013

$R_w[c(i)] = \{x : p(x)\}$

where, $p(x) = [K_d \quad [D_{c(i)}^{(x)}]] \geq cl\_\sup$

Subsequently, we find the similarity of the documents (x) with respect to the representative $D_{c(i)}$ words $R_w[c(i)]$. The definition of similarity measure plays an importance role in obtaining effective and meaningful clusters. The similarity between two text documents S_m is computed as follows,

$$S(K_d[D_{c(i)}^{(x)}], R_w[c(i)]) = \left| K_d[D_{c(i)}^{(x)}] \cap R_w[c(i)] \right|$$

$$S_m(K_d[D_{c(i)}^{(x)}], R_w[c(i)]) = \frac{S(K_d[D_{c(i)}], R_w[c(i)])}{\left| K_d[D_{c(i)}] \right| \left| R_w[c(i)] \right|}$$

The documents within the partition are sorted according to their similarity measure and a new cluster is formed when the similarity measure exceeds the minimum threshold.

IV. EXPERIMENTATION AND PERFORMANCE EVALUATION

We have implemented the proposed approach using Java (JDK 1.6). The results of the proposed approach are given in sub-section 4.1 for both the datasets: *dataset 1* (documents from different topics) and *dataset 2* (Reuter 21578 dataset). The performance of the proposed approach is evaluated on Reuter-21578 (dataset 2) [42] using F-measure. *Reuter 21578 dataset:* In 1987, the documents in the Reuters-21578 set resembled on the Reuters newswire. The documents were accumulated and indexed with grouping, by personnel from Reuters Ltd. In addition, formatting and data file production was performed in 1991 and 1992 by David D. Lewis and Peter Shoemaker at the Center for Information and Language Studies, University of Chicago.

4.1. Experimental Results

For experimentation with *dataset 1*, we take 21 documents from various topics namely, Association Rule mining (ARM) in Medical data (D_1 to D_3), Utility based ARM (D_4 to D_{10}), Biometrics (D_{11} to D_{16}) and Face recognition (D_{17} to D_{21}). Initially, the top 10 frequent words are extracted from each document and the binary database with 111 attributes is constructed. The frequent itemsets are mined from the binary database and the itemsets are sorted based on their support level. We have obtained 36 frequent itemsets of varying length from 1 to 4. Subsequently, initial partition is constructed using these frequent itemsets shown in table 1. After that, representative words of the each partition are computed based on both the top 10 and familiar words of the partition. The similarity measure (shown in table 2) is calculated for each document in the partition. The resultant cluster is formed, only if the similarity value of the documents within the partition is below 0.4. So, finally we get six clusters from four partitions shown in table 3.

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 1, Issue 7, September 2013

Table 1: Generated Partitions of text documents

Partition	Text Document	Similarity measure
P1	D4, D5, D6, D7, D8, D9, D10	0.8
P2	D11, D12, D13, D14, D15, D16	0.4
P3	D17, D18, D19, D20, D21	0.4
P1 P4	D1, D2, D3	0.6

Table 2: Similarity measure of text documents

Partition	Text Document	Similarity measure
P_1	D_4	0.8
	D_5	0.4
	D_6	0.4
	D_7	0.6
	D_8	0.6
	D_9	0.8
	D_{10}	0.6
P	D_{11}	1.0
	D_{12}	1.0
	D_{13}	0.0
	D_{14}	0.0
	D_{15}	1.0
	D_{16}	0.0
P_3	D_{17}	0.25
	D_{18}	0.25
	D_{19}	0.75
	D_{20}	0.75
	D_{21}	0.25
P_4	D_1	1.0
	D_2	0.66
	D_3	0.66

Table 3: Resultant cluster

Partition	Cluster	Text Documents
P_1	C_1	D_4, D_5, D_6, D_7, D_8, D_9, D_{10}
P 2	C_2	D_{11}, D_{12}, D_{15}
	C_3	D_{13}, D_{14}, D_{16}
P 3	C_4	D_{19}, D_{20}
	C_5	D_{17}, D_{18}, D_{21}
P_4	C_6	D_1, D_2, D_3

4.2. Performance Evaluation

We have used the following metrics namely, Precision, Recall and F-measure described in [39, 40] for evaluating the performance of the proposed approach. The evaluation metrics used in the proposed approach is given below,

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 1, Issue 7, September 2013

$$\text{Recall}(i, j) = C_{ij} / C_i$$

$$\text{Precision}(i, j) = C_{ij} / C_j$$

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

where C_{ij} is the number of members of topic i in cluster j, C_j is the number of members of cluster j and C_i is the number of members of topic i.

In order to evaluate the proposed approach on Reuter-21578 database, we have taken 125 documents from 10 different topics (cpi, bop, cocoa, coffee, crude, earn, trade, acq, money-fx, oilseed). The proposed approach uses these documents as input text and finally it results in 24 clusters. For each cluster, we compute the precision, Recall and F-measure with the help of the above mentioned equations. The obtained results are shown in table 4.

Table 4: Clustering performances obtained on Reuters-21578

Partition	Cluster	Precision	Recall	F-measure
P	C_1	0.8	0.5333	0.639976
	C_2	0.2857	0.5454	0.374975
P	C_3	1	0.5	0.666667
	C_4	0.9285	1	0.962925
P	C_5	1	0.4166	0.588169
	C_6	0.4444	0.2666	0.333269
P	C_7	1	0.2666	0.42097
	C_8	1	0.3333	0.499962
P_5	C_9	0.6666	0.1333	0.222172
P	C_{10}	1	0.2857	0.444427
	C_{11}	0.75	0.25	0.375
P	C_{12}	1	0.2857	0.444427
	C_{13}	0.5	0.0909	0.153833
P	C_{14}	1	0.2727	0.428538
	C_{15}	0.3333	0.0909	0.142843
P_9	C_{16}	1	0.0666	0.124883
P_{10}	C_{17}	0.75	0.2	0.315789
P_{11}	C_{18}	0.5	0.0833	0.142808
P	C_{19}	1	0.2727	0.428538
	C_{20}	0.6	0.2727	0.374974
P	C_{21}	1	0.5	0.666667
	C_{22}	0.5	0.25	0.333333
P_{14}	C_{23}	1	0.0714	0.133284
P_{15}	C_{24}	0.75	0.25	0.375

V. CONCLUSION

Due to the exponential increase in the volume of text document collections and the need for analyzing text documents, several techniques have been developed for mining the frequent associations from text documents. Within the text mining environment, text clustering signifies one of the most effective approaches to group documents in an unsupervised manner. In this paper, we developed an effective approach for text clustering in accordance with the frequent itemsets that provides significant dimensionality reduction. We obtained a set of non-overlapping partitions using these frequent itemsets and the resultant cluster is generated within the partition for the document collections. We used the Reuter 21578 dataset for experimentation and the clustering performance of the proposed approach was effectively analyzed.

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 1, Issue 7, September 2013

REFERENCES

[1] Hany Mahgoub, Dietmar Rosner, Nabil Ismail and Fawzy Torkey, "A Text Mining Technique Using Association Rules Extraction", *International Journal of Computational Intelligence*, Vol.4; No. 1, 2008.

[2] Shenzhi Li, Tianhao Wu, William M. Pottenger, "Distributed Higher Order Association Rule Mining Using Information Extracted from Textual Data", *ACM SIGKDD Explorations Newsletter*, Natural language processing and text mining Vol. 7, No. 1 , pp. 26 - 35 , 2005.

[3] R. Baeza-Yates, B. Ribeiro-Neto. "Modern Information Retrieval", *ACM Press*, New York,1999.

[4] J. Han, M. Kamber, "Data Mining: Concepts and Techniques", Morgan Kaufmann, SanFrancisco, 2000.

[5] Jochen Dijrre, Peter Gerstl, Roland Seiffert, "Text Mining: Finding Nuggets in Mountains of Textual Data", *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining* , San Diego, California, United States , pp: 398 - 401, 1999.

[6] Haralampos Karanikas, Christos Tjortjis and Babis Theodoulidis, "An Approach to Text Mining using Information Extraction", *Proc. Knowledge Management Theory Applications Workshop*, (KMTA 2000), Lyon, France, pp: 165-178, September 2000.

[7] Wilks Yorick, "Information Extraction as a Core Language Technology", International SummerSchool, *SCIE-97*, 1997.

[8] Ah-hwee Tan, "Text Mining: The state of the art and the challenges", *In Proceedings of thePAKDD Workshop on Knowledge Discovery from Advanced Databases*, pp. 65-70,1999.

[9] Jain, A.K., Murty, M.N., Flynn, P.J., "Data Clustering: A Review", *ACM Computing Surveys*, Vol: 31, No: 3, pp: 264-323. 1999.

[10] Feldman, R., Sanger, J., "The Text Mining Handbook", *Cambridge University Press*, 2007.

[11] Seth Grimes, "The Developing Text Mining Market", White paper from Alta PlanaCorporation, *Text Mining Summit*, 2005.

[12] M. Grobelnik, D. Mladenic, and N. Milic-Frayling, "Text Mining as Integration of SeveralRelated Research Areas: Report on KDD'2000 Workshop on Text Mining," 2000.

[13] M. Hearst, "Untangling Text Data Mining," in the *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, 1999.

[14] Alisa Kongthon, "A Text Mining Framework for Discovering Technological Intelligence to Support Science and Technology Management", Technical Report, Georgia Institute of Technology, April 2004.

[15] Brijs, Tom, Swinnen, Gilbert, Vanhoof, Koen and Wets, "Using Association Rules for Product Assortment Decisions: A Case Study", *In proceedings of Knowledge Discovery and Data Mining*, pp: 254–260, 1999.

[16] Dong, Jianning, Perrizo, William, Ding, Qin and Zhou, "The Application of Association Rule Mining to Remotely Sensed Data", *In proceedings of the ACM symposium on Applied computing*, Vol.1, pp: 340–345, 2000.

[17] Valentina Ceausu and Sylvie Despres, "Text Mining Supported Terminology Construction", *In proceedings of the 5th International Conference on Knowledge Management*, Graz, Austria,2005.

[18] Hotho, Nurnberger and Paass, "A Brief Survey of Text Mining Export", LDV Forum, Vol.20, No.2, pp.19-62, 2005.

[19] Manning and Schütze, "Foundations of statistical natural language processing", *MIT Press*,1999.

[20] Shatkay and Feldman, "Mining the Biomedical Literature in the Genomic Era: An Overview", Journal of Computational Biology, Vol.10, No.6, pp.821-855, 2003.

[21] Pegah Falinouss, "Stock Trend Prediction using News Articles", Technical Report, Lulea.University of Technology, 2007.

[22] Nasukawa and Nagano, "Text Analysis and Knowledge Mining System", *IBM Systems Journal*, Vol.40, No.4, pp.967-984, October 2001.

[23] Zhou Chong, Lu Yansheng, Zou Lei and Hu Rong, "FICW: Frequent itemset based text clustering with window constraint", *Wuhan University Journal of Natural Sciences*, Vol: 11, No: 5, pp: 1345-1351, 2006.

[24] Xiangwei Liu and Pilian He, "A Study on Text Clustering Algorithms Based on Frequent Term Sets", *Lecture Notes in Computer Science*, Vol:3584,pp:347-354, 2005.

[25] Le Wang, Li Tian, Yan Jia and Weihong Han, "A Hybrid Algorithm for Web Document Clustering Based on Frequent Term Sets and k-Means", *Lecture Notes in Computer Science, Springer Berlin* ,Vol: 4537, pp: 198-203, 2010.

[26] Zhitong Su ,Wei Song ,Manshan Lin ,Jinhong Li, "Web Text Clustering for Personalized E- learning Based on Maximal Frequent Itemsets", *Proceedings of the 2008 International Conference on Computer Science and Software Engineering* , Vol: 06, Pages: 452-455 , 2008.

[27] Yongheng Wang , Yan Jia and Shuqiang Yang, "Short Documents Clustering in Very Large Text Databases", *Lecture Notes in Computer Science, Springer Berlin* ,Vol:4256, pp: 83-93, 2006.

[28] Florian Beil, Martin Ester and Xiaowei Xu, " Frequent term-based text clustering", *in Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, Edmonton, Alberta, Canada, pp. 436 - 442 , 2002.

[29] W.-L. Liu and X.-S. Zheng, "Documents Clustering based on Frequent Term Sets", *Intelligent Systems and Control*, 2005.

[30] Henry Anaya-Sánchez, Aurora Pons-Porrata, and Rafael Berlanga-Llavori, "A document clustering algorithm for discovering and describing topics", *Pattern Recognition Letters*, Vol: 31, No: 6, pp: 502-510, April 2010.

[31] Congnan Luo, Yanjun Li and Soon M. Chung, "Text document clustering based on neighbors", *Data & Knowledge Engineering*, Vol: 68, No: 11, pp: 1271-1288, November 2009.

[32] Zamir O., Etzioni O., "Web Document Clustering: A Feasibility Demonstration", in *Proceedings of ACM SIGIR 98*, pp. 46-54, 1998.

[33] M.H.C. Law, M.A.T. Figueiredo, A.K. Jain, "Simultaneous feature selection and clustering using mixture models", *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 26(9), pp.1154-1166, 2004.

[34] Un Yong Nahm and Raymond J. Mooney, "Text mining with information extraction", *ACM*, pp. 218, 2004.

[35] Lovins, J.B. 1968: "Development of a stemming algorithm", *Mechanical Translation and*

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 1, Issue 7, September 2013

Computational Linguistics, vol. 11, pp. 22-31, 1968.

[36] Pant. G., Srinivasan. P and Menczer, F., "Crawling the Web". Web Dynamics: Adapting to Change in Content, Size, Topology and Use, edited by M. Levene and A. Poulovassilis, Springer- verilog, pp: 153-178, November 2004.

[37] R. Agrawal, T. Imielinski and A. Swami, "Mining association rules between sets of items in large databases", *In proceedings of the international Conference on Management of Data*, ACM SIGMOD, pp. 207–216, Washington, DC, May 1993.

[38] R. Agrawal and R. Srikant, "Fast algorithms for mining association rules", *In Proceedings of 20th International Conference on Very Large Data Bases*, Santiago, Chile, pp. 487–499, September 1994.

[39] Bjornar Larsen and Chinatsu Aone, "Fast and Effective Text Mining Using Linear-time Document Clustering", *in Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*, San Diego, California, United States , pp. 16 – 22,1999.

[40] Michael Steinbach, George Karypis and Vipin Kumar, "A Comparison of Document ClusteringTechniques", *in proceedings of the KDD-2000 Workshop on Text Mining*, Boston, MA, pp.109-111, 2000.

[41] B.C.M. Fung, K. Wang and M. Ester, "Hierarchical document clustering using frequent itemsets", *in Proceedings of SIAM International Conference on Data Mining*, 2003.

[42] Reuters-21578, Text Categorization Collection, UCI KDD Archive.