Frequent Item-sets Based on Document Clustering Using k-means Algorithm

Miss Jainee Patel[1], Mr. Krunal Panchal[2]

[1]*Student of Gujarat Technological University, Department of Computer Engineering, L. J. Institute of Engineering and Technology, Gujarat, India*
[2]*Assistant Professor, Department of Computer Engineering, L. J. Institute of Engineering and Technology, Gujarat, India*

ABSTRACT

The amount of text data stored in computer repositories is growing every day, we need more than ever a reliable way to assemble or classify text documents. Clustering can provide a means of introducing some form of organization to the data, which can also serve to highlight significant patterns and trends. Document clustering is used in many fields such as data mining and information retrieval. Text clustering is the method of combining text or documents which are similar and dissimilar to one another. In several text tasks, this text mining is used such as extraction of information and concept/entity, summarization of documents, modeling of relation with entity, categorization/classification and clustering. Frequent item sets for text clustering we measure the mutual overlap of frequent sets with respect to the sets of supporting documents.

Keyword - *Text Mining, Text Document, Frequent Item sets, clustering, k-means algorithm*

.

1. Introduction

It is very difficult to find the required pieces of information in the bundled of scattered information. The task becomes more challenging, when we find that over 9o% of the information available is in unstructured and semi-structured forms, which is very difficult to search. Text mining, also known as text data mining or knowledge discovery from textual databases, refers generally to the process of extracting interesting and nontrivial patterns or knowledge from unstructured text documents [3]. It can be viewed as an extension of data mining or knowledge discovery from (structured) databases.

Text mining is made up of two words: “Text” & “Mining”. It means the extraction of large volume of text to find the relevant information. Text mining is similar to data mining but it is an extended form of data mining. Data mining is extract useful information from large database.

Text mining structure comprises of two modules [3]:

A. Text refining

Text refining converts unstructured text documents into an intermediate form.

B. Knowledge distillation

In this process, Knowledge is gathered from intermediate form, which is obtained by text refining. The text mining works is that it transforms unstructured documents into Intermediate form (IF).

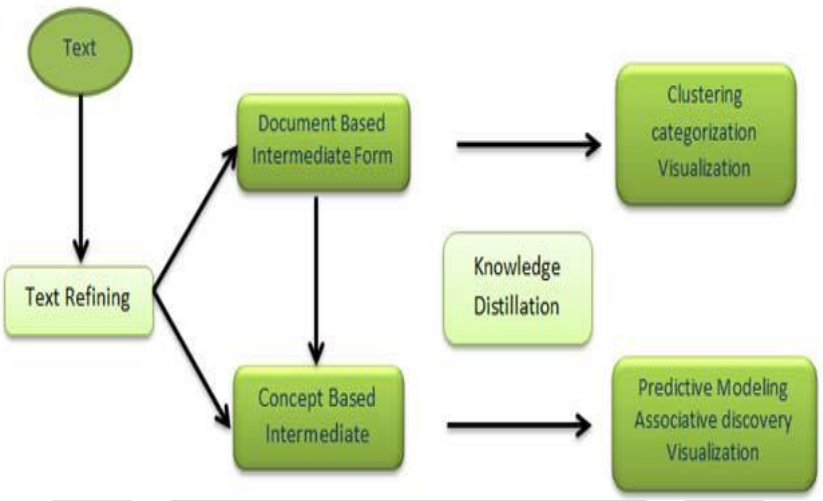

Fig. 1 A text mining structure [2]

The IF can be framed by two methods. They are [2]:

- Document based IF
- Concept based IF

Once information is converted into IF, knowledge distillation practice is executed that helps in transforming document based IF into clustered, categorized and virtualized information as well as concept based IF into predictive modeling, associate discovery or virtualized information [2].

2. MINING OF FREQUENT ITEMSETS

Frequent item sets means a Set of items that frequently appear together in a transactional data set. Frequent Item set (or pattern) Mining is acknowledged in the data mining field because of its broad applications in mining association rules, correlations, and graph pattern constraint based on frequent patterns, sequential patterns, and many other data mining tasks. Efficient algorithms for mining frequent item sets are crucial for mining association rules as well as for many other data mining tasks.

The main aim is to optimize the process of finding patterns which should be efficient, scalable and can detect the important patterns which can be used in various ways.

Frequent item-sets include two methods:

- Apriori Algorithm

Apriori is a conventional algorithm that was first introduced in for mining association rules. The two steps used for mining association rules are as follows.(1)Identifying frequent item sets (2)generating association rules from the frequent item sets. Frequent item sets can be mined in two steps. At first, candidate item sets are generated and afterwards frequent item sets are mined with the help of these candidate item sets. Frequent item sets are nothing but the item sets whose support is greater than the minimum support specified by the user [6].

- FP-Growth Algorithm

FP Growth Stands for frequent pattern growth .It is a scalable technique for mining frequent pattern in a database .Adopts divide-and-conquer strategy. It just scan database two times & use no candidate set [1]. FP-growth uses a

combination of the vertical and horizontal database layout to store the database in main memory. Instead of storing the cover for every item in the database, it stores the actual transactions from the database in a tree structure and every item has a linked list going through all transactions that contain that item. This new data structure is denoted by FP-tree (Frequent-Pattern tree) all transactions are stored in a tree data structure.

3. CATEGORIES OF CLUSTERING ALGORITHMS

The process of grouping a set of physical or abstract object into classes of similar objects is called clustering. A cluster is a collection of data object that are similar to one another with in same cluster and dissimilar to the objects in the other cluster.

Document clustering is particularly useful in many applications such as automatic categorization of documents, grouping search engine results, building taxonomy of documents, and others.

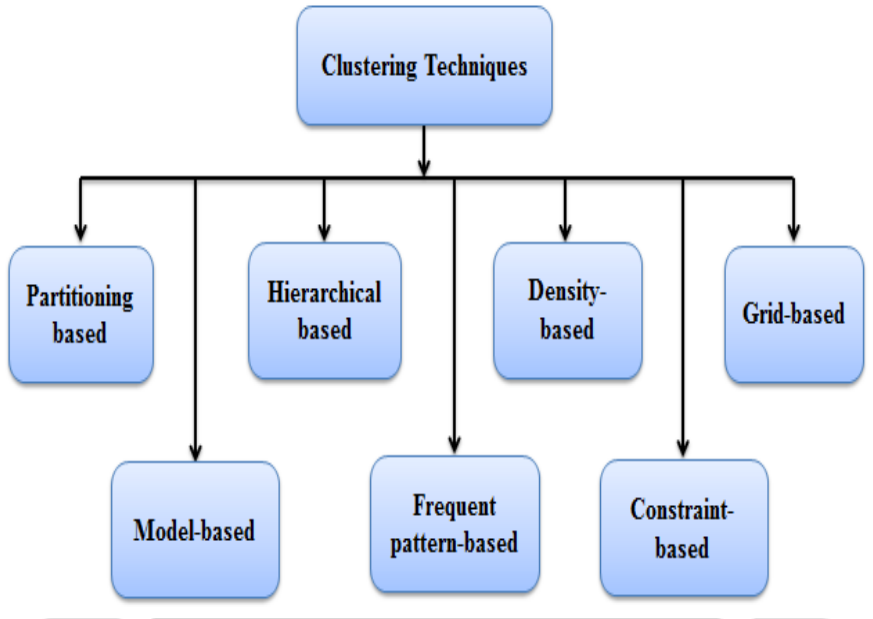

Fig. 2 classification of Document Clustering Techniques

Data clustering algorithms can be classified into following categories:

3.1 Partitioning methods

Partitional clustering algorithm creates a set of data non-overlapping subsets (clusters) such that each data object is in exactly one subset. These approaches require selecting a value for the desired number of clusters to be generated. A few popular heuristic clustering methods are k-means and a variant of k-means-bisecting k-means,k-medois,PAM,etc.

3.2 Hierarchical methods

Hierarchical clustering algorithm creates a nested set of clusters that are organized as a tree. Such hierarchical algorithms can be agglomerative or divisive approaches [1]. Agglomerative algorithms, called the bottom-up

algorithms, initially treat each object as a separate cluster and successively merge the couple of clusters that are close to one another to create new clusters until all of the clusters are merged into one [4].

Divisive algorithms called the top-down algorithms, proceed with all of the objects in the same cluster and in each successive iteration a cluster is split up using a flat clustering algorithm recursively until each object is in its own singleton cluster. The popular hierarchical methods are BIRCH, ROCK, Chameleon and UPGMA [1].

3.3 Density-based methods

Density-based clustering methods group the data objects with arbitrary shapes. Clustering is done according to a density (number of objects), density-based connectivity. The popular density-based methods are DBSCAN, OPTICS and DENCLUE [4].

3.4 Grid-based methods

Grid-based clustering methods use multi resolution grid structure to cluster the data objects. The benefit of this method is its speed in processing time. Some examples include STING, Wave Cluster [4].

3.5 Model-based methods

Model-based methods use a model for each cluster and determine the fit of the data to the given model [4]. It is also used to automatically determine the number of clusters. Examples of this methods includes Expectation-Maximization,COBWEB and SOM(self-organizing map).

3.5 Frequent pattern-based clustering

Frequent pattern-based clustering can be used in searches for patterns that occur frequently in large data sets. Frequent pattern mining can lead to the discovery of interesting associations and correlations among data objects [1].

3.6 Constraint-based clustering

Constraint-based clustering methods perform clustering based on the user-specified or application-specific constraints. It imposes user's constraints on clustering such as user's requirement or explains properties of the required clustering results [4].

4. LITERATURE REVIEW

Sr. No.	Paper Title	Method Used	Advantages	Disadvantages
1.	Frequent Patterns Mining of Stock Data Using Hybrid Clustering Association Algorithm[6]	hybrid k-mean plus MFP algorithm	Time complexity is linear which widely used, Approach is very efficient	It requires proper data with specific attributes
2.	CitiSafe: Adaptive	Fp-Growth	Privacy and	crime prevention

	Spatial Pattern Knowledge Using Fp-growth Algorithm for Crime Situation Recognition[7]	algorithm referred to as CitiSafe algorithm	security concern Software required approved authentication	and control should be Cost-effective, privacy issues
3.	Clustering digital forensic string search output[8]	k-Means, Kohonen SoM, Latent Dirichlet Allocation (LDA) followed by k-Means, LDA followed by SoM Algorithm	Improve scalability, quality measure	Required large drive space
4.	Text clustering using frequent itemsets[9]	CFWS, CMS, FTC FIHC, Maximum capturing	Better performance, improve quality, scalability	Used association rules so its required large space.
5.	A hybrid approach to automatic text summarization[1o]	KCS Approach	improve Performance, scalability	Its only measured high k-score and discarded low score text
6.	Document Clustering for Forensic Analysis: An Approach for Improving Computer Inspection[11]	(K-means, K-medoids, Single Link, Complete Link, Average Link, and CSPA)	Clustering improve performance with accuracy	Not involve cluster labelling, different methods have different accuracy

Table 1: Comparison of Literature Survey

5. The Proposed Model

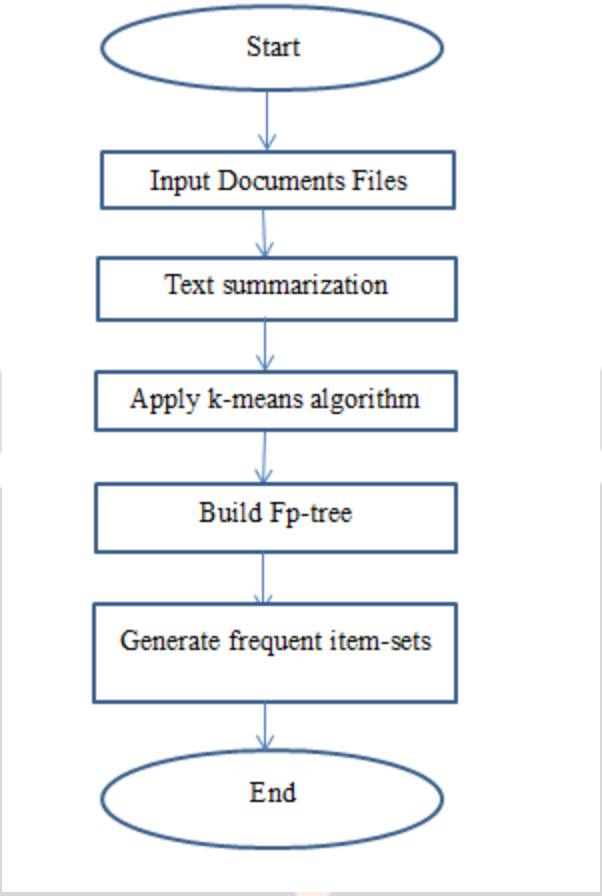

Fig.3 Flow Chart of Proposed System

Input: documents Files
Output: items are arranged frequency vise

To overcome the studied issue the Proposed Work is:

- Documents are in text format and available in large database with contain missing word, incomplete text, etc.
- Documents are contains noise data so first we will clean out.
- Clustered the all the documents by using Partitional k-means algorithms.
- FP-growth is scalable method for mining frequent pattern in a database and adopts divide-and-conquer approach.
- The proposed system will measured precision, recall

6. Implementation

```
Ten Most frequent words used with their frequency
    'WORD'          'FREQ'     'REL. FREQ'
    'of'            [  14]     '5.3640%'
    'the'           [  12]     '4.5977%'
    'to'            [   8]     '3.0651%'
    'students'      [   6]     '2.2989%'
    'The'           [   5]     '1.9157%'
    'Rs'            [   4]     '1.5326%'
    'cell'          [   3]     '1.1494%'
    '900'           [   2]     '0.7663%'
```

Fig.4 most frequent word with their frequency

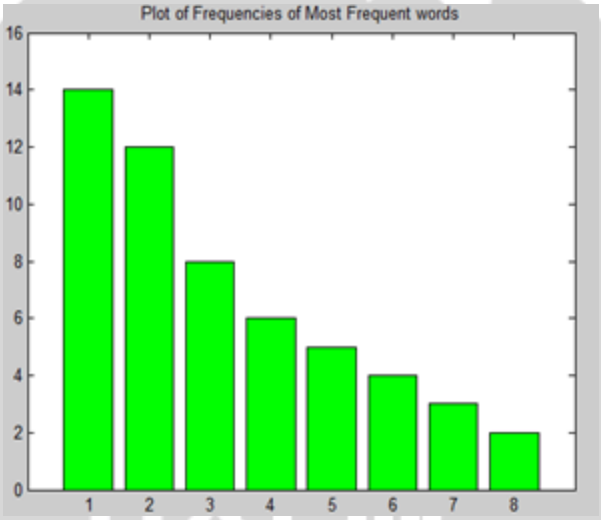

Fig.5 Plot of Most Frequent Word

Partition	Cluster	Precision	Recall
P_1	C_1	0.8	0.5333
	C_2	0.2857	0.5454
P_2	C_3	1	0.5
	C_4	0.9285	1
P_3	C_5	1	0.4166
	C_6	0.4444	0.2666
P_4	C_7	1	0.2666
	C_8	1	0.3333

7. CONCLUSIONS

Information retrieval or discover new knowledge from unstructured document is one of the main challenge in text mining. Which required clustering the document frequency wise by using some techniques? Frequent item sets mining techniques includes apriori algorithm and Fp-Growth algorithm. We compare both algorithms than find out

Fp-growth algorithm require less memory as compare to other algorithm. In this work is extract the information which documents are available in large database with contain missing word, incomplete text, etc. that are removed and arrange the document frequency wise by using some clustering techniques. The future work is we have to compared k-means algorithm and fussy k-means algorithm and generate results.

7. REFERENCES

[1] Jiawei Han,Micheline Kamber "Data Mining concept and techniques" ELSEVIER,2006

[2] Vijay Kumar Verma, Manish Ranjan, Priyanka Mishra"Text Mining and Information Professionals" IEEE,2015,DOI:978-1-4799-5532-9/15,pp.133-137

[3] A.Akilan "Text Mining: Challenges and Future Directions"IEEE Sponsered second intranational conference on electrotonics and communication systems,IEEE,2015,978-1-4788-7225-8/15,pp.1679-1684

[4] Sandeep Kumar,Associate Prof. Sanjay Pandey "Survey of Document Clustering Approach for Real World Objects (Documents)" International Journal on Recent and Innovation Trends in Computing and Communication,August 2015,ISSN:2321-8169,Volume:3 Issue: 8,pp.5073 - 5075

[5] Florian Beil,Martin Ester,Xiaowei Xu "Frequent Term-Based Text Clustering"ACM,2oo2,1-58113-567-X/02/007,PP.436-442

[6] S.Murali Krishna,S.Durga Bhavani"An Efficient Approach for Text Clustering Based on Frequent Itemsets" European Journal of Scientific Research,ISSN 1450-216X Vol.42 No.3 (2010), pp.385-396

[7] Aurangzeb Khan, Khairullah khan, Baharum B. Baharudin"Frequent Patterns Mining Of Stock Data Using Hybrid Clustering Association Algorithm"International Conference on Information Management and Engineering, IEEE, 2009, DOI:10.1109/ICIME.2009.129, pp.667-671

[8] Omowunmi E. Isafiade, Antoine B. Bagula MIEEE"CitiSafe:Adaptive Spatial Pattern Knowledge Using Fp-growth Algorithm for Crime Situation Recognition"2013 IEEE 1oth International Conference on Ubiquitous Intelligence & Computing and 2o13 IEEE 1oth International Conference on Autonomic & Trusted Computing, IEEE, 2013, DOI:10.1109/UIC-ATC.2013.72,pp.551-556

[9] Nicole L. Beebe, Lishu Liu "Clustering digital forensic string search output"Information Systems and Cyber Security Department, Elsevier, 2014, DOI:0.1016/j.diin.2014.10.002, pp.314-322

[10] Wen Zhang, Taketoshi Yoshida , Xijin Tang , Qing Wanga "Text clustering using frequent itemsets" Knowledge-Based Systems,Elsevier,2010, DOI:10.1016/j.knosys.2010.01.011, pp.379-388

[11] Te-Min Chang, Wen-Feng Hsiao A hybrid approach to automatic text summarization" IEEE, 2008,978-1-4244-2358-3/08,pp.65-70

[12] Luís Filipe da Cruz Nassif and Eduardo Raul Hruschka Document Clustering for Forensic Analysis: An Approach for Improving Computer Inspection",IEEE,2013,1556-6013,VOL.8,NO.1,pp.46-54