RESEARCH ARTICLE OPEN ACCESS

Approach by Applying Clustering Algorithm for Document Clustering

Priyanka Khadse, Harshal Chowhan
Department of Computer Science and Engineering, Nagpur University, India
Email: priyanca.khadse@gmail.com
Department of Computer Science and Engineering, Nagpur University, India
Email: chowhan23@gmail.com

ABSTRACT
Document clustering is the process of partitioning a collection of texts into subgroups including content based similar ones. The process of document clustering is used for searching information and understanding by the human. In today era, all the documents are in electronic format, because of smaller storage and quick access. It is a problem to access relevant documents from the lager dataset. Text mining is not stand-alone task that analysts typically engage in. The goal is to provide, a process for clustering the documents. The approach is basically used to extract unknown pattern from a large set of document.

***Keywords* -** Document clustering, text mining, information extraction, clustering algorithm.

I. Introduction

Document clustering is the process of dividing a collected texts into subgroups which include contents based similarities. This process is used for searching the information by the human in very less time. In other word document clustering is process which organized document automatically into meaningful clusters or group. Clustering is the process of organizing data objects into a set of different classes know as cluster. Objects that are in the same cluster are similar among themselves and different from object belong to other cluster. Clustering algorithms are typically used for exploratory data analysis, there is little or no prior knowledge about the data. This is case in many applications of document clustering from a more technical viewpoint, datasets consist of unlabeled objects—the classes or categories of documents that can be found are a priori unknown.

In this, the use of clustering algorithms, which are capable of finding latent patterns from text documents found it can enhance the analysis performed by the expert examiner.

The paper is organized as follows. Section II presents literature survey. Section III clustering algorithm and preprocessing steps. Section IV conclusion.

II. Literature Survey

The literature on document clustering only data use by algorithms which assume that the number of clusters in known and fixed a priori. Focusing on relaxing this concept which was not accepted in practical applications, a common way in various domains which involves estimating the number of clusters from data.

Document clustering is done by using different techniques and models, such as Kohonen's Self Organizing Maps (SOM) [4] and The k-means Algorithm [1]. Beebe and Dietrich in [5] proposed a new process model for text string searches that advocated the use of machine learning techniques, clustering being one of them. Clustering algorithms are typically used for exploratory data analysis, where there is little or no prior knowledge about the data [2], [3]. Document clustering has been investigated for use in a number of different areas of text mining and information retrieval. Initially, document clustering was investigated for improving the precision or recall in information retrieval systems [Rij79, Kow97] and as an efficient way of finding the nearest neighbors of a document [BL85]. More recently, clustering has been proposed for use in browsing a collection of documents [CKPT92] or in organizing the results returned by a search engine in response to a user's query [ZEMK97]. Document clustering has also been used to automatically generate hierarchical clusters of documents [KS97].

Since there are many techniques by using vector space model and preprocessing, here represent result.

III. CLUSTERING ALGORITHM AND PREPROCESSING

Clustering algorithm are used for document clustering. Agglomerative hierarchical clustering and FCM are clustering techniques that are used for document clustering in this research. Agglomerative hierarchical clustering is as "better" than K-means, although slower. FCM is work on fuzzy logic methods .

3.1. Hierarchical clustering algorithm

Hierarchical techniques produce a nested sequence of partitions, with a single, all inclusive cluster at the top and singleton clusters of individual points at the bottom. Each intermediate level can be viewed as combining two clusters from the next lower level The result of a hierarchical clustering algorithm can be graphically displayed as tree, called a dendogram. This tree graphically displays the merging process and the intermediate clusters. The dendogram at the Fig 1 shows how four points can be merged into a single cluster. For document clustering, this dendogram provides a hierarchical index. There are two basic approaches to generating a hierarchical clustering:

a) Agglomerative: Start with the points as individual clusters and, at each step, merge the most similar or closest pair of clusters. This requires a definition of cluster similarity or distance.

b) Divisive: Start with one, all-inclusive cluster and, at each step, split a cluster until only singleton clusters of individual points remain. In this case, we need to decide, at each step, which cluster to split and how to perform the split.

Agglomerative Algorithm:

1. Compute the similarity between all pairs of clusters, i.e., calculate a similarity matrix whose ijth entry gives the similarity between the ith and jth clusters.

2. Merge the most similar (closest) two clusters.

3. Update the similarity matrix to reflect the pa irwise similarity between the new cluster and the original clusters.

4. Repeat steps 2 and 3 until only a single cluster remains.

The vector space model is used to represent the frequencies of occurrence of words. The output screen in shown below in Fig 2

3.2. Preprocessing

The preprocessing is the process which remove extract word in the sentence and provide only the synonym. I have collected 21000 data of tweeter from website and blog, after collection the data, combined all that tweets into one single text file. WordNet is a lexical database for the English language. It groups English words into sets of synonyms, provides short, general definitions, and records the various semantic relations between these synonym sets. WordNet distinguishes between nouns, verbs, adjectives and adverbs because they follow different grammatical rules. It does not include prepositions, determiners etc. RiTa.wordnet is provides simple access to the wordnet. In my project I am using RiTa.wordnet for extracting the action words from the collected database. This wordnet counts the adjective and noun in the sentences, and the count of word are used in the FCM algorithm. The output of the preprocessing is shown in Fig 3, Table 1 and the screenshot.

IV. FIGURES AND TABLES

4.1. Denogram for four point.

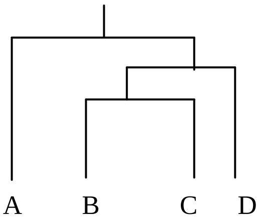
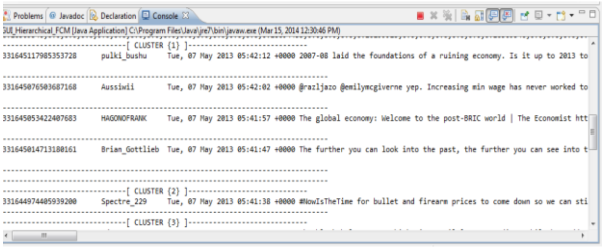


Fig.1: example of denogram .

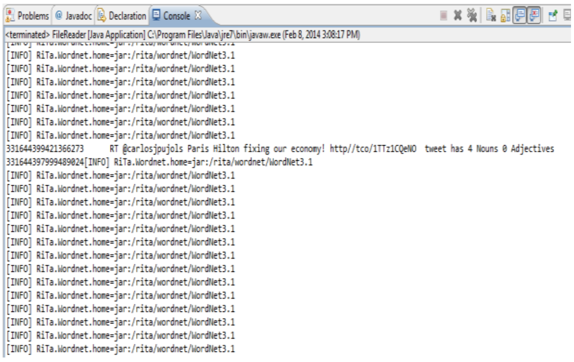
4.2. Output screen of agglomerative algorithm:



```
GUI_Hierarchical_FCM [Java Application] C:\Program Files\Java\jre7\bin\javaw.exe (Mar 15, 2014 12:30:46 PM)
----------------------------[ CLUSTER {1} ]----------------------------
331645117985353728      pulki_bushu     Tue, 07 May 2013 05:42:12 +0000 2007-08 laid the foundations of a ruining economy. Is it up to 2013 to
------------------------------------------------------------------------
331645076503687168      Aussiwii        Tue, 07 May 2013 05:42:02 +0000 @razljazo @emilymcgiverne yep. Increasing min wage has never worked to
------------------------------------------------------------------------
331645053422407683      HAGONOFRANK     Tue, 07 May 2013 05:41:57 +0000 The global economy: Welcome to the post-BRIC world | The Economist htt
------------------------------------------------------------------------
331645014713180161      Brian_Gottlieb  Tue, 07 May 2013 05:41:47 +0000 The further you can look into the past, the further you can see into t
------------------------------------------------------------------------
------------------------------------------------------------------------
----------------------------[ CLUSTER {2} ]----------------------------
331644974405939200      Spectre_229     Tue, 07 May 2013 05:41:38 +0000 #NowIsTheTime for bullet and firearm prices to come down so we can sti
------------------------------------------------------------------------
----------------------------[ CLUSTER {3} ]----------------------------
```

Fig.2: output screen of agglomerative algorithm

4.3. Preprocessing steps with example:



```
<terminated> FileReader [Java Application] C:\Program Files\Java\jre7\bin\javaw.exe (Feb 8, 2014 3:08:17 PM)
[INFO] RiTa.Wordnet.home=jar:/rita/wordnet/WordNet3.1
[INFO] RiTa.Wordnet.home=jar:/rita/wordnet/WordNet3.1
[INFO] RiTa.Wordnet.home=jar:/rita/wordnet/WordNet3.1
[INFO] RiTa.Wordnet.home=jar:/rita/wordnet/WordNet3.1
[INFO] RiTa.Wordnet.home=jar:/rita/wordnet/WordNet3.1
[INFO] RiTa.Wordnet.home=jar:/rita/wordnet/WordNet3.1
[INFO] RiTa.Wordnet.home=jar:/rita/wordnet/WordNet3.1
331644399421366273      RT @carlosjpujols Paris Hilton fixing our economy! http//tco/1TTz1CQeNO  tweet has 4 Nouns 0 Adjectives
331644397999489024[INFO] RiTa.Wordnet.home=jar:/rita/wordnet/WordNet3.1
[INFO] RiTa.Wordnet.home=jar:/rita/wordnet/WordNet3.1
[INFO] RiTa.Wordnet.home=jar:/rita/wordnet/WordNet3.1
[INFO] RiTa.Wordnet.home=jar:/rita/wordnet/WordNet3.1
[INFO] RiTa.Wordnet.home=jar:/rita/wordnet/WordNet3.1
[INFO] RiTa.Wordnet.home=jar:/rita/wordnet/WordNet3.1
[INFO] RiTa.Wordnet.home=jar:/rita/wordnet/WordNet3.1
[INFO] RiTa.Wordnet.home=jar:/rita/wordnet/WordNet3.1
[INFO] RiTa.Wordnet.home=jar:/rita/wordnet/WordNet3.1
[INFO] RiTa.Wordnet.home=jar:/rita/wordnet/WordNet3.1
[INFO] RiTa.Wordnet.home=jar:/rita/wordnet/WordNet3.1
[INFO] RiTa.Wordnet.home=jar:/rita/wordnet/WordNet3.1
[INFO] RiTa.Wordnet.home=jar:/rita/wordnet/WordNet3.1
[INFO] RiTa.Wordnet.home=jar:/rita/wordnet/WordNet3.1
[INFO] RiTa.Wordnet.home=jar:/rita/wordnet/WordNet3.1
```

Fig.3: output screen of preprocessing

4.4. Table for word net :
If the sentence are as follow then by using word net, it provide output

Table 1: WordNet Table For Preprocessing

Sentence	Synonym
Paris helton fixing our economy	4 nouns and 0 adjective
The Jaggi explains why Amartya Sen is wrong on food security Again	7 Nouns 3 Adjectives
Australia’s central bank cuts rate to 275 percent	7 Nouns 1 Adjectives

V. CONCLUSION

The study of different techniques of clustering, in this paper, I conclude that hierarchical algorithm can be used for clustering the document in predefined limitation for cluster.

FUTURE WORK

In this paper two parts hierarchical algorithm and preprocessing technique is shown with the output screen. In future by using preproceesing output, FCM will generate for document clustering and the comparison between this two algorithm based on inter and intra clustering will be done.

REFERENCES

[1] A. L. N. Fred and A. K. Jain, “Combining multiple clusterings using evidence accumulation,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 27, no. 6, pp. 835–850, Jun. 2005.

[2] TC Havens, JC Bezdek, C Leckie, LO Hall “Fuzzy c-means for very large data” ieeexplore.ieee.org - 2012

[3] A. K. Jain and R. C. Dubes, Algorithms for Clustering Data. Englewood Cliffs, NJ: Prentice-Hall, 1988.

[4] A. Strehl and J. Ghosh, “Cluster ensembles: A knowledge reuse framework for combining multiple partitions,” J. Mach. Learning Res., vol.3, pp. 583–617, 2002.

[5] X. Rui, D. Wunsch II, “Survey of Clustering Algorithms”, IEEE Transactions on Neural Networks, vol.16, no.3, 2005.