Text Mining, its Utilities, Challenges and Clustering Techniques

Bhavna Bhardwaj
M. Tech. (F.S.P.)
School of Future Studies and Planning
D.A.V.V., Indore Pin-452001, INDIA

ABSTRACT
Text analysis is an interdisciplinary field of data mining in which person try to extract meaningful results from the unstructured textual data. In this paper the focus will be on different text mining application, the problems that we face while doing text mining and different text clustering approaches and try to figure out what next can be done for better performance of clustering algorithms.

Keywords
Text mining; clustering; Opinion mining;

1. INTRODUCTION
Text mining also referred as Knowledge discovery from textual data is an interdisciplinary field of Data Mining, Statistics and Machine Learning. Basically the data mining techniques are applied on three types of data for knowledge extraction, those are – Structured data (SQL database with rows and columns), Semi-structured data (CSV files, XML files, NoSQL documents), Unstructured data (Text documents, Scientific data, Photographs and Videos).

Technically Text Mining is a process of deriving high quality information from text. In this the hidden patterns and trends are extracted from the textual data. There are multiple utilities of text mining like- Text Classification, Opinion Mining, Text Clustering, and Document Summarization etc. [1]. In this paper the discussion will be about different utilities of text mining, challenges in the field of text mining and some of the text clustering algorithms.

2. UTILITIES OF TEXT MINING
There are numerous applications of text mining those can be used for knowledge extraction purpose and can be proved very useful. Those are all described below-

2.1 Text Classification
Text classification is the technique of arranging documents into predefined groups based on their content. It is the automated assignment of natural language texts to predefined categories. There are numerous text documents available in electronic form. The task of data mining is to automatically classify documents into predefined classes based on their content. The most common techniques used for this purpose include Association Rule Mining, Implementation of Naïve Bayes Classifier, Genetic Algorithm, and Decision Tree and so on. [2]

2.2 Opinion Mining
with the explosive growth of social media on the web ,individuals and organizations are increasingly using the content in this media for their decision making .[3] In present era if a person want to buy a product, then he/she is no longer bound to ask only family or friend's view because there are numerous customer reviews available on the web. Similarly if someone want to know about the vision on anything (consumer product, any political issue, any global issue etc.) on global basis, he/she can find that by using some techniques. These are known as Opinion Mining or Sentiment Analysis.

2.3 Text Clustering
It refers to find the group of similar objects in the data set or document set. When we talk about data set it's like social media data (example- status, tweet updates) by using these we can extract important patterns regarding some specific thing. If we talk about document set then it comes to document clustering like in between 100 documents we can derive the groups of document of same domain. K-medoid , K-means etc. are some examples of clustering algorithms.

2.4 Document Summarization
Summary is a text that is derived from one or more text documents, that project important information in the original text(s) in brief.[4] The main problem in document summarization lies in recognizing the more significant parts of the document and the lesser one. This all can be done by the help of natural language processing and by measuring the frequency of words. The Naive Bayes method, Hidden Markov model and Log-linear models are some appropriate methods for summarization.[5]

3. CHALLENGES FOR TEXT MINING:
Text mining is still not that famous technique of data mining as others since it is used to extract knowledge from textual data i.e. unstructured data which is still a tough task because of some limitations those are-

- The uppermost problem in text mining is the ambiguity of the language i.e. the capability of being understood in two or more possible sense. Because one word or phrase may have multiple meanings those can lead to ambiguity problem. [6]
- In fields like Bioinformatics there are multiple names for a single gene or protein that may also lead to ambiguity problem.
- Since some filtration techniques are applied on the original data but, still sometimes some words remain in the text because they have higher frequency and affect the result.
- One more problem with test mining is when we use the social media data i.e. status updates, tweets, comments, reviews etc. most people use slang

words like- “btw” for by the way, “ppl” for people etc. these words do not exist in the dictionary that’s why they affects the mining results.

- Another problem with text mining is cleaning the data, if we extract online texts then we also get the reference addresses of the images linked with the text and those references are hard to remove.

4. CLUSTERING TECHNIQUES

Clustering is a technique of grouping similar type of objects in a data, here we are talking about textual data only that, how the clustering can be done if you have an unstructured data set. Text clustering algorithms are divided into a wide variety of diff erent types such as agglomerative clustering algorithms, partitioning algorithms, and standard parametric modelling based methods. [7]

4.1 Agglomerative And Hierarchical Clustering Algorithms

The general idea of agglomerative clustering is to progressively consolidation records into clusters in view of their comparability with each other. All the various levelled clustering calculations progressively union gatherings in view of the best pairwise comparability between these gatherings of reports. The fundamental diff erences between these classes of routines are regarding how this pairwise comparability is figured between the diff erent gatherings of reports. For instance, the comparability between a couple of gatherings may be processed as the best-case similitude, normal case likeness, or most pessimistic scenario closeness between reports which are drawn from these sets of gatherings. Reasonably, the procedure of agglomerating records into progressively larger amounts of clusters makes a cluster chain of importance (or dendogram) for which the leaf hubs relate to individual archives, and the interior hubs compare to the combined gatherings of clusters. At the point when two gatherings are combined, another hub is made in this tree comparing to this bigger consolidated gathering. The two offspring of this hub compare to the two gatherings of records which have been converged to it. [7]

The agglomerative methodologies which are used for merging groups are-

4.1.1 Single Linkage Clustering

In single linkage clustering basically we join two or more already existing pool of documents, i.e. the fundamental consideration in single linkage clustering to find the highest similarity between the pools that is the greatest similarity between any documents belong to that group. The advantage of this method is that is highly adequate for clustering purpose.[7] This algorithm is very closely related to the spanning tree algorithm.[8]

The biggest disadvantage of this algorithm is that it can lead to the anomaly of chaining for example- is X is identical to Y , and Y is identical to Z, then it doesn’t mean that X is also identical to Z i.e. Transitivity property does not applies here.[7]

4.1.2 Group-Average Linkage Clustering

Group average linkage clustering is somewhat similar to the single linkage clustering technique, here we derives the average similarity between the pairs of documents in two pools. Apparently we can understand this method is little bit slow than the previous one because of deriving the average similarity, but the algorithm is rich if we consider the quality of clusters since the chaining behaviour ends up here. The time complexity for this algorithm in $O(n^2)$ [7].

4.1.3 Complete Linkage Clustering

Complete linkage clustering method tends to overcome the shortcomings of the previous algorithm, but definitely we gets the higher complexity O(n3) here [7]. In this method initially all the documents are part of its own cluster and as the process begins, the documents steadily associates into larger cluster considering the shortest distance between them. Hence the problem of chaining gets solve here [9].

4.2 Distance-Based Partitioning Algorithms

Distance based partitioning is another approach of clustering; K-Means and K-Mediod are the two well-known algorithm of partitioning. Here the clusters are not created in a single step, numerous task are performed to make clusters and the distance calculation is the most important task in that.

4.2.1 K-Means Clustering Algorithm

The basic concept of this method is K centroids for each cluster. Initially we have to place K centroids wisely so that the distances can be calculated. The second step is to find the nearest centroid for each point to associate with it. When all the points are covered at this point we re-calculate the new centroids and after this again the binding is done on the same data points. This whole procedure takes place in a loop until we gets the minimized centroids which can’t be further change the clusters [7].

$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} \left| x_i^j - C_j \right|^2$$

4.2.2 K-Mediod Clustering Algorithm

The key aim of K-Mediod algorithm is to obtain the most utilized set of the documents represented in the whole set , each particular document is assigned to its respective cluster from the set. For this initially we select a point and then the steps of clustering work iteratively with the use of randomized procedure.[7]

The disadvantage of this method is that it needs lot of iteration which makes the process slower. Another problem is that it does not work efficiently for unstructured kind of data, for example text data, audio or video data.

5. CONCLUSION AND FUTURE WORK

Text analysis presently is really a fascinating technique to determine the useful results from the textual data. By using text mining techniques we can easily extract public reviews, can classify the text into predefined classes, can conclude the documents and also can make group or cluster of multiple documents. In this paper we have discussed few clustering algorithms there may be some other algorithm exist which may be more efficient.

Among the entire clustering algorithm discussed above the future possibility is that we can hybridize two algorithm to find a more efficient result with lesser time and space complexity. It can be done after proper compatibility testing between algorithms.

6. REFERENCES

[1] H. J. kamber and J. pei, Data Mining: Concepts and Techniques, San Francisco: Margan Kaufman, 2011.

[2] Text Classification Using Data Mining, S. M. Kamruzzaman1 Farhana Haider2 Ahmed Ryadh Hasan ICTM-2005.

[3] Web Data Mining, Chapter 9 Opinion Mining and Sentiment Analysis, Authors: Liu, Bing, Department of Computer Science, University of Illinois, Chicago, 851 S. Morgan St., Chicago, IL, 60607-7053, USA.

[4] Radev, D. R., Hovy, E., and McKeown, K. (2002). Introduction to the special issue on summarization. Computational Linguistics.

[5] A Survey on Automatic Text Summarization, Dipanjan Das Andre F.T. Martins, Language Technologies Institute Carnegie Mellon University{dipanjan, afm} @cs.cmu.edu,November 21, 2007IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 6, No 2, November 2012 ISSN (Online): 1694-0814 www.IJCSI.org.

[6] Techniques, Applications and Challenging Issue in Text Mining , Shaidah Jusoh and Hejab M. Alfawareh , College of Computer Science & Information Systems, Najran University P.O Box 1988, Najran, Saudi Arabia

[7] A SURVEY OF TEXT CLUSTERING ALGORITHMS, Charu C. Aggarwal IBM T. J. Watson Research Center Yorktown Heights, NY, charu@us.ibm.comChengXiang Zhai University of Illinois at Urbana-Champaign Urbana, IL , czhai@cs.uiuc.edu

[8] E.M. Voorhees. Implementing Agglomerative Hierarchical Clustering for use in Information Retrieval,Technical Report TR86–765, Cornell University, Ithaca, NY, July 1986.

[9] https://en.wikipedia.org/wiki/Complete-linkage_clustering