 ISSN : 0976-8491 (Online) | ISSN : 2229-4333 (Print)

The Assessment of a Text Document Clustering and Classification using Fuzzy C-Means Clustering Algorithm

[1]Shaik Sharmila, [2]Bodapati Prajna
[1,2]Dept. of Computer Science and Systems Engg., Andhra University, Visakhapatnam, AP, India

Abstract
The similarity between documents are the new creative idea now days in data mining and data recovery. These incorporate basically supported hunt, question reformulation and picture recovery. Standard text comparability measures perform ineffectively due to data meager condition and the absence of context. Where Document preparing assumes a vital part in data mining, and web look. In text handling, pack of-words model is utilized. Measuring the closeness between records is a fundamental assignment in the report preparing and text classification. In this, another comparability measure is proposed. To quantify the similitude between records regarding a component, the proposed technique takes the accompanying cases: (a) The element must be in both documents, (b) the element that shows up in one and only archive, and (c) the element that shows up in none of the documents. For first case, closeness increments as the distinction between the records highlight values diminishes. For second, a settled esteem is discover the closeness. For last case, the element has no commitment to closeness. The adequacy of measure is assessed on a few genuine data sets for record classification and clustering.

Keywords
Documents, Clustering, Clustering Analysis, Optimization, Fuzzy C-Means Clustering, Text Document Clustering

I. Introduction

Today we are confronting an always expanding volume of text archives [2]. The voluminous texts streaming over the Internet, huge gathering of archives in advanced libraries and stores, and computerized data, for example, sites and messages are growing up quickly step by step. It conveys upon the difficulties that how to sort out text documents successfully and productively. The issue of clustering has been concentrated broadly in the database and insights writing in the context of a wide assortment of data mining errands [1,5]. The clustering issue can be characterized as discovering gatherings of comparative protests in the data. The closeness between two distinct items is measured with the utilization of a similitude work. The issue of clustering can be extremely profitable in the text space, as the items to be clustered can be of various granularities, for example, archives, sections, sentences, terms and so on. Keeping in mind the end goal to apply the majority of the clustering calculations, two things are required: speaking to a protest, and a similarity measure between items. A clustering calculation finds a parcel of an arrangement of articles that satisfies some paradigm in view of comparability measure. The text clustering is the issue of naturally gathering of free text documents. The gatherings are typically depicted by an arrangement of watchwords or expressions that portrayed the normal substance of the archives in the gathering. To play out a clustering procedure, the items ought to have some sort of credits to gauge the separation or comparability among the articles. These properties are normally called as components of the question. A large portion of the recommendations in this field consider the report as an arrangement of words. In these representations, every component relates to a solitary word found in the record set. As an archive set may contain a few thousand of words, these outcomes in a high impracticable dimensionality. To diminish the report space dimensionality some word decrease strategies are connected in the pre-preparing stage. The most widely recognized technique to lessen the quantity of various words is to dispose of the words with low data esteem. These words are called stop words. The stop words are gathered into a lexicon or a rundown. Another path for decrease depends on the measurable properties of the words: the rare and the incessant words are sifted through from the first text.

II. Related Work

To compute the similarity between two records as for an element, the proposed measure considers the accompanying three cases: (a) The element shows up in both archives, (b) the component shows up in one and only report, and (c) the element shows up in none of the documents. For the main case, the similarity increments as the contrast between the two included element values diminishes. [10] Furthermore, the commitment of the distinction is ordinarily scaled. For the second case, a settled esteem is added to the comparability. For the last case, the element has no commitment to the similarity. The proposed measure is reached out to gage the closeness between two arrangements of documents [6]. Two archives are slightest like each other if none of the elements have non-zero values in both records. Furthermore, it is attractive to consider the esteem appropriation of an element for its commitment to the similitude between two records [2]. The proposed plot has additionally been stretched out to gauge the similarity between two arrangements of records. In this work, we are concentrating on the execution came about because of the utilization of various similitude measures in various classification/clustering calculations. Notion examination or feeling mining comprise of a wide range of fields like common dialect handling, text mining, basic leadership and etymology [7]. It is a sort of text examination that orders the text and settles on choice by extricating and breaking down the text. Sentiments can be classified as constructive and antagonistic and measures the level of constructive or contrary connected with that occasion, for example, individuals, association and social issues [5]. Thus, it's fundamentally individuals' feeling study, investigation of feelings and evaluations toward any social issue, individuals or substance. As of late the greater part of the inquiries about have been done on the slant investigation of items and administrations. The examination of occasions and issues, data is recovered from web-based social networking like twitter etc. [4]. The impact of utilizing unlabeled data as a part of conjunction with a little segment of marked data on the exactness of a centroid-based classifier used to perform single name text classification. To utilize centroid-based strategies since they are quick when contrasted and other classification techniques, yet at the same time introduce precision near that of the best in class strategies [1]. Productivity is especially critical for huge spaces, similar to general news bolsters, or the web.to propose the blend of Expectation-Maximization with a centroid-based technique to fuse data about the unlabelled

data amid the preparation stage [8]. To propose another option to EM in light of the incremental upgrade of a centroid-based technique with the unlabelled documents amid the preparation stage. Furthermore indicated how a centroid based strategy can be utilized to incrementally redesign the model of the data, in light of new confirmation from the unlabelled data [9].

III. Concept Overview: Clustered Based Text Classification

Conventionally clustering based classification algorithms consists of basic two steps:

A. Clustering Step

In clustering step, training data set is clustered into number of clusters, so that similar documents categorize into same cluster

B. Classification Step

In classification step, classifiers are trained by using the above formed clusters

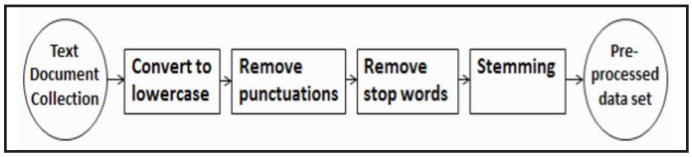

Fig. 1: Pre-processing Steps in Text Mining

Clustering of text document is one of text mining tasks. Text mining deals with unstructured data. Text documents fail to get the imposed structure of traditional database, though it out speaks a very wide range of information. Thus, it is important to represent this unstructured data into structured form, so that appropriate patterns and features can be retrieved from this text information.

A. Pre-processing of Text Dataset

The natural language pre-processing operations on text documents collection such as converting to lower case, removing punctuations and stop words, stemming and white space removal (shown in fig. 1) are required for obtaining the structured form of text data. These operations act as pre-processing task. Stop words occur most often in the text documents but they cannot make any sense in the documents.

Stop words such as a, an, the, is, at, which, on etc. are filtered out in the pre-processing of natural language data (text). Stemming in information retrieval is used to describe the process in which infected or derived words are reduced to their base or root form.

Each pre-processed document is treated as bag of words (set of all words with frequency of words appearing in that document). The term-document matrix is formed that describes the frequency of terms that occur in a collection of documents in vector space model. There are various schemes for determining the value that each entry in the matrix should take. One such scheme is TF – IDF (term frequency – inverse document frequency). They are useful in the field of natural language processing. TF– IDF is a numerical statistic which reflects how important a word is to a document in a collection or corpus. It is often used as a weighting factor in information retrieval and text mining [3]. For each term TF – IDF values are calculated and only terms that have TF – IDF value more than specific threshold are included into the vector space model. Other terms are simply neglected. This is done for dimensionality reduction of feature space.

B. Document Clustering

Clustering divides a set of documents into groups such that documents within same group are similar to each other. Documents are grouped together based on some similarity measure into different clusters. Similarity measure means similarity of content. Content based similarity is based on comparison of textual content of documents. Each document has a set of terms and associated frequencies, which help in clustering of documents. Similarity between all pairs of clusters is computed to form a similarity matrix. Documents can be clustered in many ways like hierarchical and K- means technique. Documents can be clustered into hierarchical structure suitable for browsing which suffers efficiency problems. Documents can also be clustered with k-means algorithm and its variants which are more efficient but less accurate. For using k-means clustering, documents should be represented in numeric format.

C. Cluster Classification

As clustering results can characterize the Basis for distribution of the whole data set documents, clustering is helpful to aid supervised classification of documents. Thus, clusters can be used to extract useful features and subsequently to augment training data set to improve the performance of classification.

The supervised learning of a classifier can be done using the clustered data set with sufficient features obtained so far. The benefit for integrating the clustering method in classification is that clustering methods are more robust to the bias caused by the initial sparse data [8]. Thus, clustering can be effective when data is sparse.

IV. Proposed Methodology

A. System Architecture

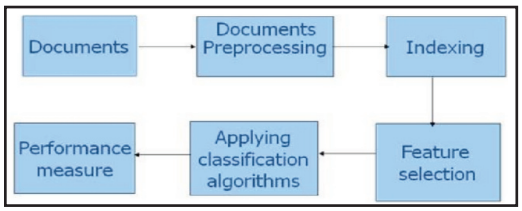

Fig. 2: Proposed System Architecture

In above fig. 2 Proposed System Architecture is given which includes several modules for analysis of similarity measures between documents and clustering process for documents.

B. Documents Representation

The bag-of-words model is a simplifying representation used in natural language processing and information retrieval (IR). In this model, a text (such as a sentence or a document) is represented as the bag of its words, disregarding grammar and even word order but keeping multiplicity. Recently, the bag-of-words model has also been used for computer vision.

- Each word corresponds to a dimension in the resulting data space
- Each document then becomes a vector consisting of non-negative values on each dimension these values are corresponds to feature
- Let D = {d1,d2,d3,d4 ....} be the set of documents & T = {t1,t2,t3,t4....tm} be the set of unique terms in D
- A document is then represented as a m dimensional vector, td = (tf(d, t1), . . . , tf(d, tm))

C. Documents Preprocessing

Many words are not informative and thus irrelevant are remove from the document representation. (e.g.) the, a, an, and, there, their, is, was, were, where, etc. These word typically about 400 to 500. It is used to improve the efficiency and potential problems of removing stop words [8].

Reducing words from their root form. A document may contain several occurrences of words like fish, fishes and fishers. An advantage of stemming is to improve the effectiveness to match similar words. Reduce indexing size to combing words with same roots may reduce indexing size as much as 40-50%. Porter algorithm is used to stem the words [8].

Porter Algorithm

Step 1: Gets rid of plurals and -ed or -ing suffixes
Step 2: Turns terminal y to i when there is another vowel in the stem
Step 3: Maps double suffixes to single ones: -ization, -ational, etc.
Step 4: Deals with suffixes, -full, -ness etc. Step 5: Takes off -ant, -ence, etc.
Step 6: Removes a final –e

Indexing

The Extracted text documents are converted into Boolean weighting by using the indexing technique of Term Frequency – Inverse Document Frequency. TF–IDF is the product of two statistics, term frequency and inverse document frequency. The term frequency tf (t, d), the simplest choice is to use the raw frequency of a term in a document, i.e. the number of times that term t occurs in document d. The raw frequency of t by f (t, d), then the simple tf scheme is tf (t, d) = f (t, d). Other possibilities include

- Boolean "frequencies": tf (t, d) = 1 if t occurs in d and 0 otherwise;
- Logarithmically scaled frequency: tf (t, d) = log (f (t, d) + 1);

The inverse document frequency is a measure of whether the term is common or rare across all documents. It is obtained by dividing the total number of documents by the number of documents containing the term, and then taking the logarithm of that quotient.

TF–IDF is calculated as tf – idf (t, d, D) = tf (t, d) × idf (t, D)

Feature Selection

- The similarity of two patterns and feature selection are the critical factors affecting the performance of classifier, especially for classification based on pattern similarity theory
- Feature selection is a technique of selecting subset of relevant features for building robust learning model

Applying Classification Algorithm

- In this randomly selected training documents are used for training/validation and the testing documents are used for testing
- The pre designated training data are used for training/ validation and the pre designated testing data are used for testing
- Note that the data for training/validation are separate from the data for testing in each case

Performance Measure

- The effectiveness of measure will be evaluated on several real-world data sets for text classification and clustering problems
- The results will show that the performance obtained by the proposed measure is better than that achieved by other measures

Proposed Algorithm

The Fuzzy c-means Algorithm

The fuzzy c-means algorithm was introduced by Ruspini and later extended by Dunn and Bezdek [9] and has been widely used in cluster analysis, pattern recognition and image processing etc. The fuzzy c-means clustering algorithm (FCM) introduces the fuzziness for the belongingness of each object and can retain more information of the data set than the hard k-means clustering algorithm (HCM). Although the FCM algorithm has considerable advantages compared to the k-means clustering algorithm, there are also some shortcomings when using the FCM algorithm in practice.

The main limitation of the FCM algorithm is its sensitivity to noises. The FCM algorithm implements the clustering task for a data set by minimizing an objective-function subject to the probabilistic constraint that the summation of all the membership degrees of every data point to all clusters must be one.

The FCM algorithm can be summarized by the following steps:
Step 1: Initialize matrix $U \in [u_{ij}]$ with the initial value U (0) ;
Step 2: At k-step: Calculate the cluster prototype matrix $V^{(k)} \in [v_i]$ with $U^{(k)}$;
Step 3: Update $U^{(k)}$, $U^{(k+1)}$;
Step 4: If $\|U^{(k+1)} - U^{(k)}\| < \varepsilon$ then stop, or to step 2.

V. Results and Discussion

The results have been obtained for two Reuters data sets. The word count results that are obtained for first dataset is given below.

Word	Term Frequency	Inverse Frequency
'said'	105	2.0950%
'mln'	75	1.4964%
'dlr'	61	1.2171%
'reuter'	58	1.1572%
'vs'	51	1.0176%
'ct'	48	0.9577%
'feb'	45	0.8978%
'-'	43	0.8579%
'26'	37	0.7383%
'compani'	36	0.7183%

The word count results that are obtained for second dataset is given below.

Word	Term Frequency	Inverse Frequency
'said'	126	2.4677%
'march'	83	1.6255%
'mln'	81	1.5864%
'-'	74	1.4493%
'3'	70	1.3709%
'pct'	68	1.3318%
'vs'	64	1.2534%
'dlr'	52	1.0184%
'compani'	43	0.8421%
'ct'	36	0.7051%

The similarity between the documents is 0.2770
The results of general count and FCM count has been simulated for two datasets and the final output is shown below in fig. 3 and fig. 4.

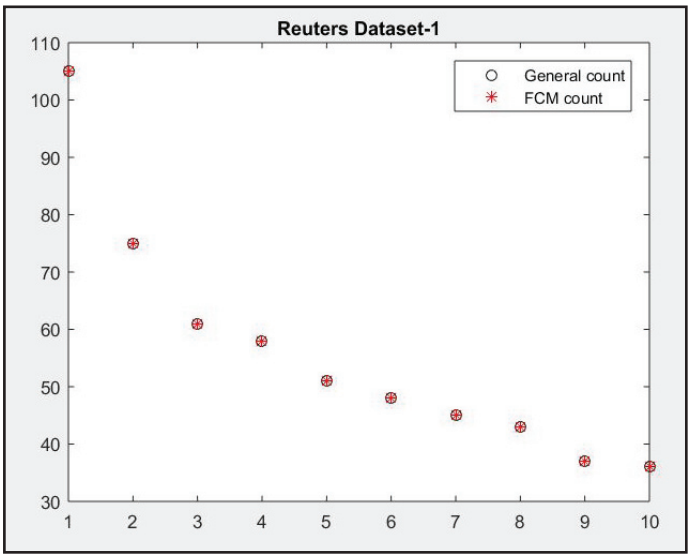

Fig 3: General and FCM count for Reuters DS1

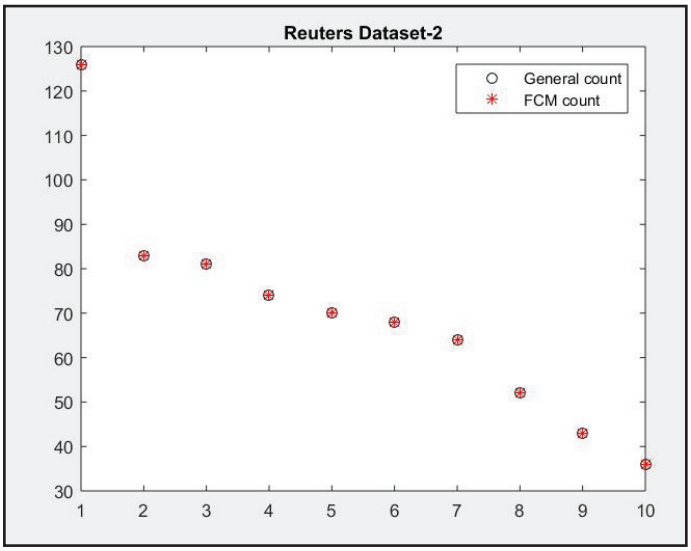

Fig. 4: General and FCM count for Reuters DS2

VI. Conclusion

Classification problems on text data mainly focus on feature space and relationship between features and classes. This paper presented a brief review on clustering based classification techniques. The central theme in many of these is providing dimensionality reduction to improve text document classification. Clustering helps in reduction of the number of redundant features, which subsequently help in reducing the dimensions. A very important goal is to achieve high quality information from text available on web. This high quality information helps in clustering and classification. Clustering on text data usually requires- First, parsing that converts unstructured to structured text. Second, text pre-processing operations is to be performed on collection of structured data to obtain pre-processed data. Third, pattern and feature extraction and also similarity measure calculations on text data by mining the knowledge. Fourth, efficient technique for clustering is to be applied to form the clusters with similar documents.

References

[1] A. Benghabrit, B. Ouhbi, H. Behja, B. Frikh,"Text clustering using statistical and semantic data", Proceedings of the IEEE World Congress on Computer and Information Technology, Jun. 22-24, pp. 1-6, 2013.

[2] Andrew Skabar, Khaled Abdalgader,"Clustering Sentence Level Text Using a Novel Fuzzy Relational Clustering Algorithm", IEEE Transactions on Knowledge and Data Engineering. 25: pp. 62-75. 2013.

[3] Prof. B. Prajna, Prof. Shashi M,"Document Clustering Technique based on Noun Hypernyms", IJECT Vol. 2, SP-1, Dec. 2011.

[4] B. Drakshayani, E. V. Prasad,"Text Document Clustering based on Semantics. International Journal of Computer Applications, 45: pp. 7-12, 2012.

[5] Nadempalli Sneha, B. Prajna, Sharmila Sujatha, "Application for Retriving Details of Users - Topic Based Approach", pp. 509-513, Vol. 6, Issue 8, IJCSET, 2015.

[6] Beil, M. Ester, X. Xu,"Frequent term-based text clustering", Proceedings of the ACM KDD Conference. USA, pp. 436-442, 2002.

[7] C. Salton, Buckley,"Term Weighting Approaches in Automatic Text Retrieval", Information Processing and Management. 24, pp. 513–523, 1988.

[8] K. Naga Neerja, B.Prajna,"An effective Research Paper Recommender System based on Subspace Clustering", International Journal Of Engineering And Computer Science, pp. 13306-13310, Vol. 4, Issue 7, July 2015.

[9] Charu C. Aggarwal, ChengXiang Zhai,"A Survey of Text Clustering Algorithms. Mining Text Data", Springer US. pp: 77-128, 2012.

[10] Daniel Zlacky, Jan Stas, Jozef Juhar, Anton Cizmar, "Slovak Text Document Clustering", Acta Electrotechnica et Informatica, 13, pp. 3–7, 2013.

[11] Shady Shehata, Fakhri Karray, Mohamed S. Kamel, "An Efficient Concept-Based Mining Model for Enhancing Text Clustering", IEEE Transactions on Knowledge and Data Engineering, Vol. 22, No. 10, October 2010.

[12] D. Renukadevi, S. Sumathi,"Term Based Similarity Measure for Text Classification and Clustering using Fuzzy c-means algorithm", International Journal of Science, Engineering and Technology Research (IJSETR), Vol. 3, Issue 4, April 2014.

[13] Ms.Gaurangi Patil, Ms. Varsha Galande, Ms. Kalpana Dange,"Sentiment Analysis Using Support Vector Machine", International Journal of Innovative Research in Computer and Communication Engineering, Vol. 2, Issue 1, January 2014.

ISSN : 0976-8491 (Online) | ISSN : 2229-4333 (Print)

Shaik Sharmila is pursuing M.Tech (CST-Artifial Intelligence and Robotics) in the Dept of Computer Science and Systems Engineering, Andhra University, Visakhapatnam, Andhra Pradesh, India. She obtained her B.Tech (CSE) from Gayatri Vidya Parishad college of Engineering, Visakhapatnam.

Bodapati Prajna, M.Tech, Ph.D is working as a Professor in Dept. of Computer Science and Systems Engineering, Andhra University, Visakhapatnam, Andhra Pradesh, India. Her research field is Data Mining.