Information Sciences 311 (2015) 149–162

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

A similarity assessment technique for effective grouping of documents

Tanmay Basu *, C.A. Murthy

Machine Intelligence Unit, Indian Statistical Institute, Kolkata 700108, India

ARTICLE INFO

Article history:
Received 20 February 2014
Received in revised form 25 December 2014
Accepted 15 March 2015
Available online 21 March 2015

Keywords:
Document clustering
Text mining
Applied data mining

ABSTRACT

Document clustering refers to the task of grouping similar documents and segregating dissimilar documents. It is very useful to find meaningful categories from a large corpus. In practice, the task to categorize a corpus is not so easy, since it generally contains huge documents and the document vectors are high dimensional. This paper introduces a hybrid document clustering technique by combining a new hierarchical and the traditional k-means clustering techniques. A distance function is proposed to find the distance between the hierarchical clusters. Initially the algorithm constructs some clusters by the hierarchical clustering technique using the new distance function. Then *k*-means algorithm is performed by using the centroids of the hierarchical clusters to group the documents that are not included in the hierarchical clusters. The major advantage of the proposed distance function is that it is able to find the nature of the corpora by varying a similarity threshold. Thus the proposed clustering technique does not require the number of clusters prior to executing the algorithm. In this way the initial random selection of *k* centroids for *k*-means algorithm is not needed for the proposed method. The experimental evaluation using Reuter, Ohsumed and various TREC data sets shows that the proposed method performs significantly better than several other document clustering techniques. *F*-measure and normalized mutual information are used to show that the proposed method is effectively grouping the text data sets.

© 2015 Elsevier Inc. All rights reserved.

1. Introduction

Clustering algorithms partition a data set into several groups such that the data points in the same group are close to each other and the points across groups are far from each other [9]. The document clustering algorithms try to identify inherent grouping of the documents to produce good quality clusters for text data sets. In recent years it has been recognized that partitional clustering algorithms e.g., *k*-means, buckshot are advantageous due to their low computational complexity. On the other hand these algorithms need the knowledge of the number of clusters. Generally document corpora are huge in size with high dimensionality. Hence it is not so easy to estimate the number of clusters for any real life document corpus. Hierarchical clustering techniques do not need the knowledge of number of clusters, but a stopping criterion is needed to terminate the algorithms. Finding a specific stopping criterion is difficult for large data sets.

The main difficulty of most of the document clustering techniques is to determine the (content) similarity of a pair of documents for putting them into the same cluster [3]. Generally cosine similarity is used to determine the content similarity between two documents [24]. Cosine similarity actually checks the number of common terms present in the documents. If

* Corresponding author. Tel.: +91 33 25753109; fax: +91 33 25783357.
E-mail addresses: mailtanmaybasu@gmail.com (T. Basu), murthy@isical.ac.in (C.A. Murthy).

http://dx.doi.org/10.1016/j.ins.2015.03.038
0020-0255/© 2015 Elsevier Inc. All rights reserved.

two documents contain many common terms then they are very likely to be similar. The difficulty is that there is no clear explanation as to how many common terms can identify two documents as similar. The text data sets are high dimensional data set and most of the terms do not occur in each document. Hence the issue is to find the content similarity in such a way so that it can restrict the low similarity values. The actual content similarity between two documents may not be found properly by checking only the individual terms of the documents. A new distance function is proposed to find distance between two clusters based on a similarity measure, extensive similarity between documents. Intuitively, the extensive similarity restricts the low (content) similarity values by a predefined threshold and then determines the similarity between two documents by finding their distance with every other document in the corpus. It assigns a score to each pair of documents to measure the degree of content similarity. A threshold is set on the content similarity value of the document vectors to restrict the low similarity values. A histogram thresholding based method is used to estimate the value of the threshold from the similarity matrix of a corpus.

A new hybrid document clustering algorithm is proposed, which is a combination of a hierarchical and k-means clustering technique. The hierarchical clustering technique produces some baseline clusters by using the proposed cluster distance function. The hierarchical clusters are named as *baseline clusters*. These clusters are created in such a way that the documents inside a cluster are very similar to each other. Actually the extensive similarity of all pair of documents of a baseline cluster is very high. The documents of two different baseline clusters are very dissimilar to each other. Thus the baseline clusters intuitively determine the actual categories of the document collection. Generally there exist some singleton clusters after constructing the hierarchical clusters. The distance between a singleton cluster and each baseline cluster is not so small. Hence k-means clustering algorithm is performed to group these documents to a particular baseline cluster, with which it has highest content similarity. If for several iterations of k-means algorithm each of these singleton clusters are grouped to the same baseline cluster then they are likely to be assigned correctly. The significant property of the proposed technique is that it can automatically identify the number of clusters. It has become clear from the experiments that the number of clusters of each corpus is very close to the actual category. The experimental analysis using several well known TREC and Reuter data sets have shown that the proposed method performs significantly better than several existing document clustering algorithms.

The paper is organized as follows. Section 2 describes some related works. The document representation technique is presented in Section 3. The proposed document clustering technique is explained in Section 4. The evaluation criteria for evaluating the clusters generated by a particular method is described in Section 5. Section 6 presents the experimental results and a detailed analysis on the results. Finally we conclude and discuss about the further scope of this work in Section 7.

2. Related works

There are two basic types of document clustering techniques available in the literature – *hierarchical* and *partitional* clustering techniques [8,11].

Hierarchical clustering produces a hierarchical tree of clusters where each individual level can be viewed as a combination of clusters in the next lower level. This hierarchical structure of clusters is also known as dendrogram. The hierarchical clustering techniques can be divided into two parts – *agglomerative* and *divisive*. In an *Agglomerative Hierarchical Clustering* (AHC) method [30], starting with each document as individual cluster, at each step, the most similar clusters are merged until a given termination condition is satisfied. In a divisive method, starting with the whole set of documents as a single cluster, the method splits a cluster into smaller clusters at each step until a given termination condition is satisfied. Several halting criteria for AHC algorithms have been proposed. But no widely acceptable halting criterion is available for these algorithms. As a result some good clusters may be merged, which will be eventually meaningless to the user. There are mainly three variations of AHC techniques – single-link, complete-link and group-average hierarchical method for document clustering [6].

In *single-link* method the similarity between a pair of clusters is calculated as the similarity between the two most similar documents where each document represents each individual cluster. The complete-link method measures the similarity between a pair of clusters as the least similar documents, one of which is in each cluster. The group average method merges two clusters if they have least average similarity than the other clusters. Average similarity means the average of the similarities between the documents of each cluster. In a *divisive hierarchical clustering* technique, initially, the method assumes the whole data set as a single cluster. Then at each step, the method chooses one of the existing clusters and splits it into two. The process continues till only singleton clusters remain or it reaches a given halting criterion. Generally the cluster with the least overall similarity is chosen for splitting [30].

In a recent study, Lai et al. have proposed an agglomerative hierarchical clustering algorithm by using dynamic k-nearest neighbor list for each cluster. The clustering technique is named as *Dynamic k-Nearest Neighbor Algorithm* (DKNNA) [16]. The method uses a list of dynamic k nearest neighbors to store k nearest neighbors of each cluster. Initially the method assumes each document as a cluster and finds the k nearest neighbors of each cluster. The minimum distant clusters are merged and their nearest neighbors are updated accordingly and then again finds the minimum distant clusters and merge them and so on. The algorithm continues until the desired number of clusters are obtained. In the merging and updating process of each iteration, the k nearest neighbors of the clusters, which are affected by the merging process are updated. If the set of k nearest neighbors are empty for some of the clusters being updated, their nearest neighbors are determined by searching all the clusters. Thus the proposed approach can guarantee the exactness of the nearest neighbors of a cluster and can obtain good quality clusters [16]. Although the algorithm has shown good results for some artificial and image data sets, but it has two

Download English Version:

https://daneshyari.com/en/article/393254

Download Persian Version:

https://daneshyari.com/article/393254

Daneshyari.com