An improved Similarity Measure For Chinese Text Clustering

Shaolei Zhang, Zhong Wang, Wei Huang
Xi'an Research Institute of High Technology, Xi'an, China
jianjianyuanqu11@163.com

Keywords: similarity measure; part of speech; initial cluster center; LDA; TF-IDF

Abstract. Similarity measure between documents is a pivotal step in text processing filed. Traditional similarity just considers one aspect of the text feature. A new similarity measure proposed in this paper takes statistics information and part of speech of feature terms into account. The proportion of statistics information and semantic, importance of different part of speech are obtained through experiment. K-means algorithm and its variants are widely used for text clustering, especially in large dataset. The choice of initial cluster centers is important, which can affect iterations and cluster quality. We proposed a new method based on previous researches. The method selects initial cluster center by combining maximum distance and statistical features. The experiments show that the improved method improves cluster quality in terms of F-measure, and has a less time consumption.

1 Introduction

Clustering is unsupervised learning technique through which objects are classified into groups called clusters. The goal is to group these objects such that objects in the same clusters should be as similar as possible and objects in the different clusters should be as dissimilar as possible[1][2]. Document clustering is used to organize text documents without prior information, which is different from text classification. K-means algorithm, classical flat clustering algorithm[3], is most widely used, which is relatively expandable and high efficiency. However, the K-means algorithm has its limitations: need to give prospective cluster K and random selection of initial cluster centers.

Several methods of dealing with the K-means algorithm suboptimality also exist: literature [4] proposes a classical improved algorithm K-means++, the algorithm can find the initial clustering center widely distributed and get better clustering results. However the cluster center is strongly random, which leads to the instable clustering result and poor reproducibility. An algorithm for initialization of K-Means clustering center based on optimized division [5] is proposed. This new algorithm could divide the data sample space optimized with histogram method, and identify the initial cluster centers obeying the natural character of data space. But it is low-efficiency when it comes to high-dimensional data object. Literature [6] presents an initial clustering centers selection algorithm based on LDA model for the K-means algorithm. The algorithm improves the precision of clustering, but the complexity of initial clustering center increases the time consumption. To solve the problem that the traditional k-means algorithm has sensitivity to the initial start center, Literature [7] presents a new method of finding the initial cluster center by computing the density of the area where the data object belongs to, whereas it is difficult to determine density parameters, which has a great influence on performance.

Accurate clustering requires a precise definition of the closeness between a pair of objects, in terms of either the pairwised similarity or distance. A variety of similarity (distance) measures have been proposed and widely applied, for example, cosine similarity and the Jaccard correlation coefficient. These classical similarity measures are commonly based on vector space model (VSM), lack of text semantic information. We propose a new measure for similarity between two documents, and several characteristics are embedded in this measure. The term frequency and semantic information are considered, and the contribution of part of speech is considered to reduce the time consumption. The proposed measure and other classical similarity are compared in the experiment, in which classical K-means and an improved k-means are clustering algorithm.
2 Text Preprocessing

Text document is weakly structured or semi-structured object. To archive text clustering, it is essential to model text documents. Vector space model (VSM) and Latent Dirichlet Allocation are two common methods using bag of words model.

2.1 Vector Space model

Vector space model is an algebraic model for representing text documents as vectors of identifiers, which is commonly used in NLP. In this model, each Text $d$ is considered as a vector in a vector space: $d = \{(t_1, w_1), (t_2, w_2), \ldots, (t_n, w_n)\}$. TF-IDF is used as a measure of characteristic vector in this paper, and this measure gives the weight of each word $t$. See Formula 1 for the calculation of the weight:

$$tf-idf(d, t) = tf(dt) \cdot \log \frac{N}{df(t)}$$

(1)

2.2 Latent Dirichlet Allocation (LDA)

LDA is a generative probabilistic model [8][9], including a three-level structure with word, topic and document. Fig 1 is the graphical model representation. In LDA, documents are viewed as a distribution over topics while each topic is a distribution over words. To generate a document, LDA firstly samples a document-specific multinomial distribution over topics from a Dirichlet distribution. Then it repeatedly samples the words from these topics. LDA and its variants have been successfully applied in many works. Fig 2 is the graphical model representation of LDA.

![Figure 1. The topology of LDA implicit theme structure.](image)

![Figure 2. Graphical model representation of LDA.](image)

3 Similarity measures

There is no measure that is universally best for all kinds of clustering problems. Choosing an appropriate similarity measure is crucial for cluster analysis. According to literature [10], experiments which compare the effectiveness of measures have been conducted. In this paper, cosine similarity and JS (Jensen-Shannon) distance is used to measure the similarity between the texts.

3.1 Cosine Similarity

When documents are represented as term vectors, the similarity of two documents corresponds to the correlation between the vectors. This is quantified as the cosine of the angle between vectors, that is, the so-called cosine similarity. Cosine similarity is one of the most popular similarity measure applied to text documents and define the similarity of two texts $d_1$ and $d_2$ in formula:

$$S_{cos}(d_1, d_2) = \frac{d_1 \cdot d_2}{\|d_1\| \|d_2\|}$$

(2)

3.2 JS (Jensen-Shannon) Distance

LDA model use topic probability vector express text, so traditional similarity such as Euclidean Distance is not suitable to compute similarity in LDA model. The Jensen–Shannon divergence[11] is a popular method of measuring the similarity between two probability distributions. The definition of JS distance is

$$d_{js} = \frac{1}{2} \left( \sum_{j=1}^{M} p_j \ln \frac{p_j}{q_j} + \sum_{j=1}^{M} q_j \ln \frac{q_j}{p_j} \right)$$

(3)

3.3 Proposed Similarity Measure

Similarity measures based on TF-IDF do not take in account semantic information of text, and the accuracy of result is not high, even mistakes [12]. Implicit semantic information is mined in topic model, while conversion from high dimension feature space to low dimension feature space leads to that text information is incomplete. So in this article we combine VSM model and LDA model to make up the drawbacks of two methods. Considering the contribution of different parts of speech in
the text is different in text clustering, we divide the words into different word sets according to the part of speech and model them respectively. We would integrate these models according to certain proportion. We would compute text similarity as formula:

\[ S(i, j) = \alpha S_{\text{cos}}(d_i, d_j) + (1 - \alpha)(1 - d_{i,j}) \]  

(4)

\[ S_{\text{cos}}(d_i, d_j) = \sum_k S_{ij}(d_{ik}, d_{jk}) \]

(5)

\[ d_{i,j} = \sum_k d_{ik}, d_{jk} \]

4 Optimizing Initial Cluster Center

K-means clustering algorithm clusters data into predefined number of clusters, and starts with initialization of cluster centroid randomly then assign data object to the most similar cluster. This process ends when a termination criterion (or predefined number of iteration have been done) meets.

K-means algorithm is sensitive to the initial cluster centers during the clustering. An optimized initial clustering center is proposed in this paper to find a set of data which can reflect the characteristics of data distribution. The procedure of initial cluster center is:

(1) Compute the distance between each two points of the dataset, and choose two objects whose distance is maximum as the two initial cluster centers.

(2) In remain \( N - 2 \) data objects, select other cluster center according to the formula:

\[ \text{Dis}(d_{i1}, d_{i2}) \times \text{Dis}(d_{j1}, d_{j2}) \geq \text{Dis}(d_{i1}, d_j) \times \text{Dis}(d_{j1}, d_j) \]

(6)

where \( \text{Dis}(d_{i1}, d_{i2}) \) is the distance between two objects.

(3) The cluster center should follow statistical characteristics to avoid noise data and isolated data.

5 Experiments

In this section, we investigate the effectiveness of our proposed similarity measure at first. The investigation is done by compare it with other similarity measures. To test the effectiveness of the improved algorithm, the original algorithms and the improved algorithms are compared.

5.1 Data And Evaluation

We choose data from the corpus of Chinese text classification (www.sogou.com/labs/dl/ c.html) as experiment dataset. The dataset consists of 6000 documents, included in 10 classes such as Finance, IT, Education etc. In order to know the performance for efficiency and quality of clustering, we employ F-measure and running time as our evaluation measure.

5.2 Results and Analysis

The proportion of noun, verb and other word set are obtained through experiment based on data set which are separate from the testing data. To compare the performance of similarity measures, we choose traditional k-means algorithm and improved algorithm as clustering algorithm. The table 1 show the cluster result varies with datasets from different similarity measures.

<table>
<thead>
<tr>
<th>Similarity measure</th>
<th>Original K-means Algorithm</th>
<th>Improved Algorithm</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cosine</td>
<td>0.723</td>
<td>0.814</td>
</tr>
<tr>
<td>Jensen-Shannon</td>
<td>0.686</td>
<td>0.769</td>
</tr>
<tr>
<td>The proposed similarity</td>
<td>0.788</td>
<td>0.883</td>
</tr>
</tbody>
</table>
Table 1 show the F-measure varies from similarity measure. From the cluster result, The proposed similarity have a better cluster quality under the F-measure criteria. As shown in Figure 5, the time consumption of the improved cluster algorithm is less than original k-means algorithm.

6 Conclusions

We have proposed a novel similarity measure between two documents, improving clustering quality by combining text statistics information, text semantic information and part of speech. We have investigated the efficiency of the proposed measure by comparison with others. The result have shown that the proposed measure have a better effect on cluster accuracy and time consumption. Besides, we proposed a method to optimize the selection of initial clustering centers, which improves the stability of clustering promoting the evaluation metrics of F-Measure. Accuracy and time speed are two aspects of cluster analysis, especially in the time of rapid growth of data. The next research is to increase the speed of clustering.

References