Comparative Study of Clustering Techniques for Short Text Documents

Aniket Rangrej [*]
raniket@cse.iitm.ac.in

Sayali Kulkarni
sayali.kulkarni@gmail.com

Ashish V. Tendulkar
ashishvt@cse.iitm.ac.in

ABSTRACT

We compare various document clustering techniques including K-means, SVD-based method and a graph-based approach and their performance on short text data collected from Twitter. We define a measure for evaluating the cluster error with these techniques. Observations show that graph-based approach using affinity propagation performs best in clustering short text data with minimal cluster error.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Clustering

General Terms

Algorithms, Experimentation

Keywords

clustering, short text, K-means, SVD, affinity propagation

1. INTRODUCTION

Document clustering has been studied for quite a while and has wide applications like search result grouping and categorization. It also forms a base for applications like topic extraction and content filtering. Techniques based on K-means and hierarchical agglomerative clustering are well-known for the same.

With the increasing popularity of micro-blogging and social networking, the documents collected from the web are becoming more and more condensed. Such data imposes new challenges in applying pristine document clustering techniques on them due to the sparseness. Clustering such microblogs and discussions could lead to new trends in web search and other such applications. In this paper, we present a comparative study of various techniques that can be applied for clustering such short text documents based on their performance.

2. SHORT TEXT CLUSTERING

The major challenge in handling short text documents is to deal with the sparsity of the words in them. Typically, documents are represented as a TFIDF feature vectors, where a document represents a data point in d-dimensional space where d is the size of the corpus vocabulary. Document doc_i is represented as $(v_1, v_2, ...v_d)$ where v_j is the TFIDF for of j^{th} word in doc_i.

In case of short text, since the term frequency of most of the words is limited in our documents (mostly 1, rarely 2 or 3) the TFIDF vector would actually boil down to a pure IDF vector. It may even be sufficient to represent the document as a 1/0 vector depending on the presence/absence of a word. In the following sections we discuss use of different clustering techniques on short text data.

2.1 K-means clustering

We begin with K-means clustering. The important factors in K-means is defining the distance measure between two data points and defining the number of clusters. We used two variations of distance measures: one is derived from cosine based similarity and the other is derived from Jaccard similarity coefficient. These are defined as below:

$$dist_{cos}(doc_i, doc_j) = 1 - \frac{\sum_{k=1}^{d} doc_i^{(k)} \times doc_j^{(k)}}{||doc_i|| \times ||doc_j||} \quad (1)$$

$$dist_{jac}(doc_i, doc_j) = 1 - \frac{|doc_i \cap doc_j|}{|doc_i \cup doc_j|} \quad (2)$$

where
$||doc_i|| = \sqrt{\sum_{k=1}^{d}(doc_i^{(k)})^2}$,
$|doc_i \cup doc_j| = \#$ distinct words either in doc_i or in doc_j and
$|doc_i \cap doc_j| = \#$ common words in both doc_i and doc_j.

We varied the number of clusters manually and observed the effect of this on the performance and choose the number with minimum error.

2.2 Singular value decomposition

Singular value decomposition and its relatives like LSI and PCA can be used in topic identification of documents [2]. Using SVD, a $m \times n$ matrix, say X, is factored as: $X = U\Sigma V^T$ where U is $m \times t$ matrix, V^T is $t \times n$ matrix, and Σ is a diagonal matrix of $t \times t$.

Here, we define matrix X as $[doc_i]$ with one row per document, where X is $n \times d$ where n is the number of documents and d is the vocabulary size. This term-document matrix decomposes into: topic-document (U), topic-topic similarity (Σ) and term-topic (V^T). The topic-document matrix (U) is of importance to us since it represents the association between a document and a topic using which we identify the most prevalent topic in the document. Documents with same topic will lie in the same cluster. Hence the documents, which are highly associated with the same topic, are clustered together [6].

Though unlikely in short texts like tweets, a document can have more than one topics. So we consider two variations based on *overlapping factor*: 1) identify a single most prevalent topic per document (1-overlap) and 2) identify top two most prevalent topics (2-overlap). In later, we consider both the topics for a document when clustering and hence a document may overlap in multiple clusters.

[*] Part of this work was done when the author was working with Persistent Systems Limited, India

WWW 2011, March 28–April 1, 2011, Hyderabad, India.
ACM 978-1-4503-0637-9/11/03.

	K-means	SVD	AffineProp
Cluster-error	6.61%	7.47%	2.95%

Table 1: Affinity propagation based method works best for short text document clustering

2.3 Affinity Propagation

Graph based algorithms are used widely in document processing. Here, the corpus is defined as a graph $G = (N, E)$ with N nodes, each node representing a document and E edges. The edges represent the similarity among the documents corresponding to the nodes that it connects. We consider a complete graph. Edge weight for edge e connecting node doc_i and doc_j is defined as $w_{ij} = sim(doc_i, doc_j)$. We use the similarity measures based on the two distance measures defined in section 2.1. We use affinity propagation on this graph to get the clusters [3].

3. EXPERIMENTS

3.1 Data Set

Twitter[1] readily provides a large corpus of short text in the form of tweets. Each tweet represents one document. TweetMotif [5] summarizes the tweeter data and enables us to search for different topics to get tweets from that topic. We handpicked 611 tweets from Twitter using the TweetMotif from different topics like programming language, computer networks, cricket, astronomy. The cleaned data currently has 1678 distinct words after removal of stop words.

3.2 Preprocessing

This includes removing the stop words[2] and stemming the words to their base form using Porter Stemmer. Each cleaned document is then converted to an IDF vector or 1/0 bit vector as required and used for further processing.

3.3 Evaluation

Most of the cluster evaluation techniques are based on cluster density [7]. However, in our setting we need to define how close the clusters are as compared to our golden set. We define a metric for clustering error based on this as below.

First we define a cluster similarity matrix, CM, for similarity between documents based on the clusters to which they belong. CM is $n \times n$ where each cell, $CM(i, j) = 1$ if doc_i and doc_j are in the same cluster and 0 otherwise. We construct two such matrices, one for gold standard clusters, CM_m and the other, CM_a, for the clustered discovered using the proposed method. The clustering error is now defined as:

$$cluster\ error = \frac{\#\ of\ 1s\ in\ (CM_m \oplus CM_a)}{n \times (n-1)/2} \qquad (3)$$

where $A \oplus B$ is element-wise XOR of matrix A and B.

Using this measure, the graph based method using affinity propagation gives the least error of 2.95%. The comparison of two proposed distance measures is given in table 2. Jaccard based measure does better than cosine based measure, especially in K-means. This could be specifically due to the characteristics of the documents imposed by the limited lengths. Effect of multiple topics in a single document is summarized in Table 3 by considering the two variations in the SVD based approach. As seen, 2-overlap results in more error due to the noise that is getting added due to the top two topics being considered for every document. This is inline to what is expected since the tweets are short enough and belong to a single topic only.

	Cosine-based	Jaccard-based
K-means	10.25%	6.61%
Affinity Propagation	2.95%	3.29%

Table 2: Comparison of different distance measures

	1-overlap	2-overlap
Cluster-error	7.47%	32.26%

Table 3: Effect of overlapping factor in SVD-based approach

4. CONCLUDING REMARKS

We studied various clustering techniques on short text documents and provided experimental results using micro-blogs corpus from Twitter. Then we defined an evaluation measure for studying the effectiveness of each of the clustering algorithms. We plan to extend our work in short text clustering to a larger scale and also improvise them by trying to extract more signal even though the data is sparse. We believe that adding information from external sources, exploiting characteristics of the tweets themselves and feature enrichment can add value in achieving the same.

5. REFERENCES

[1] Somnath Banerjee, Krishnan Ramanathan, and Ajay Gupta, *Clustering short texts using wikipedia*, SIGIR '07: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval (New York, NY, USA), ACM, 2007, pp. 787–788.

[2] Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman, *Indexing by latent semantic analysis*, Journal of the American Society for Information Science **41** (1990), 391–407.

[3] Brendan J. Frey and Delbert Dueck, *Clustering by passing messages between data points*, Science **315** (2007), 972–976.

[4] Jeon hyung Kang, Kristina Lerman, and Plangprasopchok Anon , *Analyzing microblogs with affinity propagation*, Proceedings of KDD workshop on Social Media Analytic, July 2010.

[5] Brendan O'Connor, Michel Krieger, and David Ahn, *Tweetmotif: Exploratory search and topic summarization for twitter*, ICWSM, 2010.

[6] Nordianah Ab Samat, Masrah Azrifah Azmi Murad, Muhamad Taufik Abdullah, and Rodziah Atan, *Malay documents clustering algorithm based on singular value decomposition*.

[7] M. Steinbach, G. Karypis, and V. Kumar, *A comparison of document clustering techniques*, Technical Report 00-034, University of Minnesota, 2000.

[1] http://twitter.com/

[2] http://drupal.org/files/issues/stopwords.patch