Measuring Similarity Between Documents

Mohit Sharma-11434
Pranjal Singh-10511
Indian Institute of Technology,Kanpur

October 21, 2013

1 Introduction

We aim at clustering documents from different domains and experiment with different simlilarity measures and compare their performances on different datasets. Clustering is an important and vivid technique, in general that organizes a group with a substantial number of data objects into a much smaller number of coherent groups.

2 Motivation

We have a large number of text documents to deal with and this number is increasing day by day. The abundant texts flowing over the Internet, huge collections of documents in digital libraries and repositories, and digitized personal information such as blog articles and emails are piling up quickly everyday. This calls for effective and efficient organization of text documents. These kind of clustering provides a structure which facilitates various future NLP tasks on these documents.

3 Methodology

The first step will be to model the text document. We will model it as bag of words. Words are counted in the bag, which differs from the mathematical definition of set. Each word corresponds to a dimension in the resulting data space and each document then becomes a vector consisting of non-negative values on each dimension. We will use frequency of words as weight making high frequent words more important. Before modelling, we need to preprocess the dataset. So we will first remove the *stopwords* which have almost negligible contribution. Secondly, we will stem the words using *Porter Stemmer*. We will also set a threshold for the frequency of a word so that words below that frequency will be neglected.

We will use standard *k-means* algorithm to do the clustering. So before making the clusters, a similarity/distance measure must be determined. We will use following similarity measures:

- Euclidean Distance
- Cosine Similarity
- Jaccard Coefficient
- Pearson Correlation Coefficient
- Averaged Kullback-Leibler Divergence

We will use *Purity* and *Entropy* for evaluation.The purity measure evaluates the coherence of a cluster, that is, the degree to which a cluster contains documents from a single category. Given a particular cluster C_i of size n_i, the purity of C_i is formally defined as:

$$P(C_i) = \frac{1}{n_i} max_h(n_i^h)$$

where $max_h(n_i^h)$ is the number of documents that are from i the dominant category in cluster C_i and n_i^h represents the number of documents from cluster C_i assigned to category h.
The entropy measure evaluates the distribution of categories in a given cluster. The entropy of a cluster C_i with size n_i is defined to be:

$$E(C_i) = -\frac{1}{logc} \sum_{h=1}^{k} \frac{n_i^h}{n_i} log(\frac{n_i^h}{n_i})$$

where c is the total number of categories in the data set and n_i^h is the number of documents from the h^{th} class that were assigned to cluster C_i.

4 Dataset

We will use following datasets:

- 20news
- classic
- hitech
- re0
- tr41
- wap
- webkb

References

[1] Ramiz M Aliguliyev. Performance evaluation of density-based clustering methods. *Information Sciences*, 179(20):3583–3602, 2009.

[2] Anna Huang. Similarity measures for text document clustering. In *Proceedings of the Sixth New Zealand Computer Science Research Student Conference (NZCSRSC2008), Christchurch, New Zealand*, pages 49–56, 2008.