Document Clustering: Similarity Measures

Shouvik Sachdeva (11693)
Bhupendra Kastore (11204)
Indian Institute of Technology, Kanpur

February 28, 2014

1 Introduction

Document clustering is a method to classify the documents into a small number of coherent groups or clusters by using appropriate similarity measures. We will try to test which similarity measure performs the best across various domains of text articles in English and Hindi.

2 Motivation

Document clustering plays a vital role in document organization, topic extraction and information retrieval. With the ever increasing number of high dimensional datasets over the internet, the need for efficient clustering algorithms has risen. A lot of these documents share a large proportion of lexically equivalent terms. We will exploit this feature by using a "bag of words" model to represent the content of a document.

3 Approach

In order to be able to apply any similarity measures to cluster text we need to represent the document in an appropriate manner. A convenient model to represent a document is the bag of words model, commonly used in data mining and information retrieval. The vector space model maps each document into a vector consisting of non-negative values on each dimension. Each dimension corresponds to a separate term and its value or weight depends on its frequency. We will use the tf-idf scheme to assign weights to each term as high frequency words are neither necessarily descriptive nor necessarily the most important ones. First we will have to preprocess the document removing the words which have almost negligible contribution to the corpus. We will then use the Porter Stemmer to stem the words and neglect certain words with frequency below a threshold.
After preprocessing, we will cluster the document using the k-means clustering technique. Now, to measure the degree of closeness between cluster objects in the k-means algorithm we need a similarity measure. We will cluster the documents using the following similarity measures:

- Cosine Distance
- Jaccard Coefficient
- Euclidean Distance

- Pearson Correlation Coefficient
- Averaged Kullback-Leibler Divergence

For evaluation we will use purity and entropy measures. Entropy measures the distribution of categories in a given cluster. The entropy of a cluster C_i with size n_i is defined as

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

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 this cluster C_i.
Purity provides provides insight into the coherence of a cluster i.e. the degree to which a cluster contains documents from a single category. Purity for a given cluster C_i of size n_i is given by:

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

where $max_h(n_i^h)$ is the number of documents that are from the dominant category incluster C_i and n_i^h represents the number of documents from the cluster assigned to category h.

4 Datasets

We will use the following datasets for our project:

- 20news
- BBC/BBC Sport
- Wikipedia
- FIRE

References

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

[2] N. Sandhya. Analysis of Similarity measures for text clustering. *CSC Journals, Volume 2, GRIET,* 2010.