Document Clustering: Similarity Measures

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

April 30, 2014

Abstract

Document clustering is a technique to classify a given set of documents into a certain number of groups based on some notion of closeness between the documents. The internet contains a large number of high dimensional data which needs to be classified on some grounds to enable efficient processing and organization of data. For instance, blogs, E-commerce sites, social networking sites, etc. use various clustering techniques for this purpose. Clustering techniques exploit the fact that most of the documents of a particular class contain similar kinds of words and frequency of these words can be used to predict which class that document might belong to. A model that does exactly that is the "Bag of Words model. We will use this model to represent each document and then compare these representations on the basis of various notions of similarity. We will try to test which similarity measure performs the best across various domains of text articles in English and Hindi.

1 Introduction

Owing to its wide range of applications over the internet, text clustering has becoming an interesting research problem. Classifying documents requires the notion of similarity between any two documents. Without any prior knowledge about the documents, there is no fixed notion as to how close any two documents are. In our project, we try to find out what measure of similarity works best for various kinds of documents. Documents used in the project are from varied sources in English and Hindi. The similarity measures compared are Euclidean Distance, Cosine Distance, Jaccard Distance, Pearson Correlation Coefficient, Manhattan Distance and Chebychev Distance. Manhattan Distance and Chebychev distances are not very widely used in clustering as they don't perform as well in most situations. We chose to test these measures as well to look into the possibility of abstracting certain information about our clustering. We also tried to improve upon the existing pre-processing techniques for Hindi documents.

We will first represent our document using the bag of words and the vector space model. Then we will cluster documents (now high dimensional vectors) by k-means clustering techniques using different similarity measures. Documents we will use are from varied domains from English and Hindi. We will then compare the performance of each similarity measure across the different kinds of documents. Entropy and Purity measure will be used for the purposes of evaluation.

2 Methodology

2.1 Document Representation

2.1.1 Bag of Words: Model

In this model, each word is assumed to be independent and the order in which they occur is immaterial. Each unique word is the same as another dimension in the new vector space and the component of a vector along this dimension is the frequency of the word. Hence, we can represent each document as this vector with each component containing the frequencies on each dimension.

2.1.2 Representing the document formally

We represent the document as an m-dimensional vector $\vec{t_d}$,

$$\vec{t_d} = (tf(d, t_1), \ldots, tf(d, t_m))$$

where $tf(d, t)$ denotes the frequency of the term $t \in T$ in document $d \in D$.

2.1.3 Pre-processing

First, we will remove stop words (non-descriptive such as a, and, are and do). We will use the one implemented in the python NLTK. Second, words will be stemmed using Porter's Stemmer, to map words with similar roots to a single word. For example automates, automatic, automation, automate will be mapped to the stem automat. For Hindi, due to the lack of a decent stemmer, we created a stemmer with the help of the Hindi WordNet. We computed the top list of suffixes and prefixes and used these lists for the purpose of stemming. Third, we selected the top 2000 unique words ranked by their weights and use them in our experiments. This was done to remove the contribution of low frequency words. PCA was used to reduce the dimension of the vector each document is represented by.

2.1.4 TFIDF

Now, some words might occur frequently in some documents but their frequency is relatively very low in most other documents. These words help us classify those group of documents in which their frequency is high as similar to each other. That is why we will use $tfidf$ (term frequency and inversed document frequency) instead of just the term frequency to elevate the effect of such words.

$Tfidf$ is defined as:

$$tfidf(d, t) = tf(d, t) \times log(\frac{|D|}{df(t)})$$

Where d is a document and t is the term and $df(t)$ is the number of documents in which term t appears.

2.2 Similarity Measures

A metric d is defined on a set A if it satisfies the following properties $\forall x, y, z \in A$:
$d : A \times A \rightarrow R$

1. $d(x, y) \geq 0$
2. $d(x, y) = d(y, x)$

3. $d(x, z) \leq d(x, y) + d(y, z)$

4. $d(x, y) = 0$ iff $x = y$

2.2.1 Euclidean Distance

It is the L2 norm widely used in geometric problems. The Euclidean distance is defined as

$$D_E(\vec{t_a}, \vec{t_b}) = (\sum_{t=1}^{m} |w_{t,a} - w_{t,b}|^2)^{1/2}$$

where t denotes the term and w denotes the $tfidf$.

2.2.2 Cosine Similarity

Cosine similarity is a measure of similarity between two vectors that measures the angle between them. The Cosine Distance is defined as:

$$SIM_C(\vec{t_a}, \vec{t_b}) = \frac{\vec{t_a} \cdot \vec{t_b}}{|\vec{t_a}| \times |\vec{t_b}|}$$

Its value belongs to the interval $[0, 1]$.

2.2.3 Jaccard Coefficient

The Jaccard coefficient measures similarity between two finite sets to capture the number of elements common among the two sets. The Jaccard Coefficient is defined as

$$SIM_J(\vec{t_a}, \vec{t_b}) = \frac{\vec{t_a} \cdot \vec{t_b}}{|\vec{t_a}|^2 + |\vec{t_b}|^2 - \vec{t_a} \cdot \vec{t_b}}$$

Its value belongs to the interval $[0, 1]$.

2.2.4 Pearson Correlation Coefficient

The Pearson Correlation coefficient is a measure that captures the linear dependence between two vectors. It assigns a value between -1 and +1. The Pearson Correlation Coefficient is defined as:

$$SIM_P(\vec{t_a}, \vec{t_b}) = \frac{m \sum_{t=1}^{m} w_{t,a} \times w_{t,b} - TF_a \times TF_b}{\sqrt{[m \sum_{t=1}^{m} w_{t,a}^2 - TF_a^2][m \sum_{t=1}^{m} w_{t,b}^2 - TF_b^2]}}$$

2.2.5 Manhattan Distance

The Manhattan Distance is the distance between two points if we only move along the axes of the space hat contains the two points. The Manhattan Distance is defined as

$$SIM_M(\vec{t_a}, \vec{t_b}) = \sum_{t=1}^{m} |w_{t,a} - w_{t,b}|$$

where t denotes the term and w denotes the $tfidf$.

2.2.6 Chebychev Distance

The Chebychev distance between two points is the maximum distance between the points in any single dimension. The Chebychev Distance is defined as

$$SIM_{Ch}(\vec{t_a}, \vec{t_b}) = \max_t |w_{t,a} - w_{t,b}|$$

where t denotes the term and w denotes the $tfidf$.

3 Clustering Algorithm

k-means Algorithm involves clustering the given data into k groups based on the distance between the observation points and the cluster centroids. Clusters can be initialized randomly by any point from the observation data. For best results, while choosing these random points, choose points as far away from each other and the chosen points. Different values of k should be tested before choosing the best fit.

After the clusters have been initialized, points are assigned a cluster based on the closeness of that point to the cluster. After all points have been assigned a cluster, the cluster centroids are updated and the points are reassigned to their new clusters. k needs to be chosen effectively. We have chosen $k = 3$ for our experiments.

3.1 Evaluation

3.1.1 Entropy

Entropy is a measure that provides insight into the variation of the types of documents in a given cluster. The entropy of a cluster C_i with size n_i is defined as

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

where cluster C_i contains n_i^h number of documents corresponding to that class.

3.1.2 Purity

Purity tells us to what extent a given cluster contains the documents form the dominant class. Purity for a given cluster C_i of size n_i is given by:

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

Here the maximum over all classes is taken to consider only the number of documents corresponding to the dominant class.

3.2 Datasets

We used the following datasets for our project:

Data	Description	No. of Documents	No. of Classes
20news	News Articles	19997	20
reuters	News Posts	1504	4
webkb	Online documents	4196	4
classic	Academic documents	7089	4
hindi-1	News Articles	100	4
hindi-2	Novels	17	4
7sectors	Web Pages	4556	7

Table 1: Datasets

We created two datasets Hindi-1 and Hindi-2 for comparing similarity measures in Hindi. Hindi-1 consists of news articles from four different categories namely, sports, business, politics and environment. Hindi-2 is a collection of novels and poems from four different authors.

4 Related Work

The following two tables contain entropy and purity results from a paper by Anna Huang.

Data	Euclidean	Cosine	Jaccard	Pearson	KLD
20news	0.1	0.5	0.5	0.5	0.38
classic	0.56	0.85	0.98	0.85	0.84
hitech	0.29	0.54	0.51	0.56	0.53
re0	0.53	0.78	0.75	0.78	0.77
tr41	0.71	0.71	0.72	0.78	0.64
wap	0.32	0.62	0.63	0.61	0.61
webkb	0.42	0.68	0.57	0.67	0.75

Table 2: Purity Results

Data	Euclidean	Cosine	Jaccard	Pearson	KLD
20news	0.95	0.49	0.51	0.49	0.54
classic	0.78	0.29	0.06	0.27	0.3
hitech	0.92	0.64	0.68	0.65	0.63
re0	0.6	0.27	0.33	0.26	0.25
tr41	0.62	0.33	0.34	0.3	0.38
wap	0.75	0.39	0.4	0.39	0.4
webkb	0.93	0.6	0.74	0.61	0.51

Table 3: Entropy Results

4.1 Our Results

The following are the best entropy and purity results obtained after running the k-means algorithm for 30 iterations with $k = 3$.

Data	Euclidean	Cosine	Pearson	Jaccard	Manhattan	Chebychev
20news	0.1	0.45	0.5	0.48	0.5	0.52
reuters	0.45	0.70	0.75	0.70	0.75	0.78
webkb	0.5	0.57	0.65	0.55	0.6	0.6
classic	0.6	0.65	0.8	0.75	0.8	0.75
hindi-1	0.32	0.4	0.36	0.40	0.43	0.34
hindi-2	0.51	0.56	0.57	0.52	0.58	0.51
7sectors	0.5	0.75	0.70	0.80	0.78	0.85

Table 4: Purity Results

Data	Euclidean	Cosine	Pearson	Jaccard	Manhattan	Chebychev
20news	0.80	0.55	0.5	0.62	0.55	0.50
reuters	0.55	0.50	0.30	0.40	0.28	0.35
webkb	0.80	0.65	0.55	0.80	0.52	0.60
classic	0.70	0.40	0.30	0.20	0.25	0.18
hindi-1	0.33	0.73	0.60	0.75	0.93	0.85
hindi-2	0.73	0.90	0.75	0.88	0.9	0.88
7sectors	0.65	0.43	0.38	0.43	0.30	0.40

Table 5: Entropy Results

5 Conclusions

We did clustering for various English and Hindi datasets using various similarity measures. The quality of clustering depends on the similarity measure chosen, the clustering algorithm used and the construction of the tfidf matrix. There is no similarity measure which gave best results on every data set but in general, Euclidean distance performed poorly whereas cosine and jaccard distances did fairly well on most datasets. We also observed that there is a difference between the performance of these measures in Hindi and English. For Hindi, euclidean gave better results compared to English datasets whereas cosine performed relatively poorly.

6 Future Work

These results can be improved upon by looking into more efficient preprocessing of the document by using other methods of dimensionality reduction and a better corpus for stop words. For Hindi, the stemmer results can be improved by using a practically obtained set of prefixes and suffixes. We inferred that Manhattan and Chebychev distances gave comparable results to more popular metrics. Therefore, new distance metrics can be explored and experimented with by introducing certain heuristics while computing these distances.

7 Acknowledgments

We would like to thank Professor Amitabha Mukherjee for his regular guidance and support for the project. We would also like to thank Shashwat Chandra for providing us with a basic hindi stemmer and Srijan R Shetty for providing us with the Hindi-2 dataset.

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] Ana Cardoso-Cachopo Improving Methods for Single-label Text Categorization, PhD Thesis, October, 2007.

[3] Paul S. Bradley and Usama M. Fayyad Refining Initial Points for K-Means Clustering. *In Proceedings of the 15th International Conference on Machine Learning (ICML98),* 1998.

[4] Kiri Wagstaff, Claire Cardie, Seth Rogers, Stefan Schroedl Constrained K-means Clustering with Background Knowledge. *Proceedings of the Eighteenth International Conference on Machine Learning, pages* 577 − 584*,* 2001.