Document Clustering: Similarity Measures

Bhupendra Kastore[1] , Shouvik Sachdeva[1]

[1]Indian Institute of Technology, Kanpur

Abstract

Document clustering is a method to classify the documents into a small number of coherent groups or clusters by using appropriate similarity measures. 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. We will group “similar" documents together to form a coherent cluster. This “similarity" can be defined in various ways. In the vector space, it is closely related to the notion of distance which can be defined in several ways. We will try to test which similarity measure performs the best across various domains of text articles in English and Hindi.

Similarity Measures

The various similarity measures used during clustering are:

Cosine Similarity: It is a measure of similarity between two vectors of an inner product space that measures the cosine of the angle between them. Given two documents the Cosine similarity is defined as

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

Jaccard coefficient: It measures similarity between finite sample sets, and is defined as the size of the intersection divided by the size of the union of the sample sets. Given two documents the Jaccard Coefficient is defined as

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

Pearson Correlation coefficient : It is a measure of the linear correlation (dependence) between two variables X and Y, giving a value between +1 and -1 inclusive, where 1 is total positive correlation, 0 is no correlation, and -1 is total negative correlation. Given two documents the Pearson Correlation Coefficient is defined as

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

Manhattan Distance: It is the distance that would be traveled to get from one data point to the other if a grid-like path is followed. The Manhattan distance between two items is the sum of the differences of their corresponding components. Given two documents, the Manhattan Distance between them is defined as

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

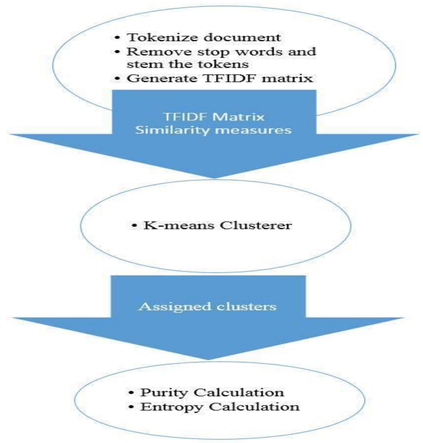
Chebychev distance: It is defined as the maximum distance between the points in any single Dimension. Given two documents the Chebychev Distance is defined as

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

Euclidean Distance: Given two documents the Eucledian Distance is defined as

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

Methodology



We are using ‘Bag of Words’ model, in which each word is assumed to be independent and the order of its occurrence is not important. Each word corresponds to a dimension in resulting data space. Each document can then be represented by a vector consisting on non-negative values in each dimension. In preprocessing phase, we first tokenize the document, then we remove all the stop words and then stemmer operates in this set of tokens. After this, we take all the unique tokens left and take a set of high frequency tokens from them, over which a document is converted into vector. We then find the TFIDF matrix of the set of document. This matrix is feed into K-means clusterer which returns a cluster value assigned to each document. We also need to mention a similarity measure on the basis of which distance between two documents is computed required during clustering.
The quality of clustering is usually measured in terms of

Purity:

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

Entropy:

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

Results

Past Results

Purity

DATA	EUCLIDEAN	COSINE	PEARSON	JACCARD	KLD
20 newsgroup	0.1	0.5	0.5	0.5	0.38
Re0	0.53	0.78	0.78	0.75	0.77
Webkb	0.42	0.68	0.67	0.57	0.75
Classic	0.56	0.85	0.85	0.98	0.84
WAP	0.32	0.62	0.61	0.63	0.61
Hitech	0.29	0.54	0.56	0.51	0.53
Tr41	0.71	0.71	0.78	0.72	0.64

Entropy

DATA	EUCLIDEAN	COSINE	PEARSON	JACCARD	KLD
20 newsgroup	0.95	0.49	0.49	0.51	0.54
Re0	0.6	0.27	0.26	0.33	0.25
Webkb	0.93	0.6	0.61	0.74	0.51
Classic	0.78	0.29	0.27	0.06	0.3
WAP	0.75	0.39	0.39	0.4	0.4
Hitech	0.92	0.64	0.65	0.68	0.63
Tr41	0.62	0.33	0.3	0.34	0.38

Our Results

Purity

DATA	EUCLIDEAN	COSINE	PEARSON	JACCARD	MANHATTAN	CHEBYCHEV
20 newsgroup	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	0.60	0.65	0.8	0.75	0.80	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

Entropy

DATA	EUCLIDEAN	COSINE	PEARSON	JACCARD	MANHATTAN	CHEBYCHEV
20 newsgroup	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	0.9	0.75	0.88	0.9	0.88
7Sectors	0.65	0.43	0.38	0.43	0.30	0.40

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.

Future Improvements

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.

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, 2001, p. 577{584.

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 improved hindi stemmer and Srijan R Shetty for providing us with a hindi dataset.