Text Clustering using Semantics

Bhoopesh Choudhary
CSE Department
Indian Institute of Technology, Bombay
India.
bhoopesh@cse.iitb.ac.in

Pushpak Bhattacharyya[*]
CSE Department
Indian Institute of Technology, Bombay
India
pb@cse.iitb.ac.in

Abstract

In traditional document clustering methods, a document is considered a bag of words. The fact that the words may be semantically related- a crucial information for clustering- is not taken into account. In this paper we describe a new method for generating feature vectors, using the semantic relations between the words in a sentence. The semantic relations are captured by the Universal Networking Language (UNL), which is a recently proposed semantic representation for sentences. The clustering method applied to the feature vectors is the Kohonen Self Organizing Maps (SOM). This is a neural network based technique, which takes the vectors as inputs and forms a document map in which similar documents are mapped to the same or nearby neurons. Experiments show that if we use the UNL method for feature vector generation, clustering tends to perform better than when the term frequency based method is used.

Keywords: Text clustering, Document vectors, Semantic net/graph, Universal Networking language, Self Organization Maps.

Topic of the paper: Text Clustering using meaning.

1 Introduction

There are many algorithms for automatic clustering like the *K Means algorithm* [Hartigan and Wong 1979], *Expectation Maximization* [Dempster et. al.1977] and *hierarchical clustering* [Jain and Dubes, 1988] which can be applied to a set of vectors to form the clusters. Traditionally the document is represented by the frequency of the words that make up the document (the Vector space model and the Self-organizing semantic map [T. Kohonen, 1995]). Different words are then given importance according to different criteria like *Inverse Document frequency* and *Information Gain*. A comparative evaluation of feature selection methods for text documents can be found in [Yang and Pedersen 1997]. These methods consider the document as a bag of words, and do not exploit the relations that may exist between the words.

However, this can cause problems. For example, if we consider the two sentences *John eats the apple standing beside the tree* and *The apple tree stands beside John's house.* On the other hand there may be some sentences, which have the same meaning but have been constructed from different sets of words. For example in the sentences, *John is an intelligent boy* and *John is a brilliant lad*, mean more or less the same thing. There are some methods like Latent Semantic Indexing [Deerwester et. al. 1995], which try to solve it. The word category map method can also be used for the same purpose. Another shortcoming of these methods is due to *polysemy* or *homography* where a word has different meanings or meaning shades in different contexts (for example, the word *bank* in *He went to the bank to withdraw some money* and *The boat was beside the bank).* It has been shown [Gonzalo et. al. 1998] that if we index words with their wordnet synset or sense then it improves the information retrieval performance.

In this paper we describe a **new method for the creation of document vectors**. This approach uses the Universal Networking Language (UNL) representation of a document. The UNL (explained in detail in [Uchida, Zhu and Della 1995]) represents the document in the form of a semantic graph with universal words as nodes and the semantic relation between them as links. For example, figure 1 shows the UNL graph for the sentence *John, who is the chairman of the company, has arranged a meeting at his residence.* Instead of considering the documents as a bag of words we use the information given by the UNL graph to construct the vector.

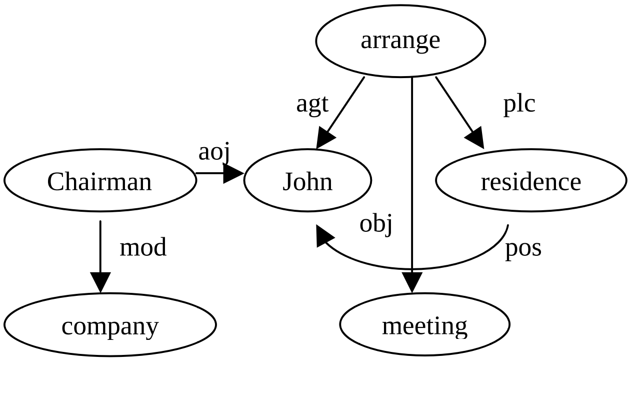


Figure 1: An example UNL graph

[*] Contacting author

2 Document Vector Construction Using UNL Graph Links

In the UNL link method, instead of using the words as components for the document vector we use the Universal Words- which are concepts formed using English words and attaching restrictions to them- as the components of the vector. Since each UW is disambiguated (for example the financial bank is represented as *bank (icl>financial institute)* and the river bank is represented as *bank (mod>river)* in UNL), multiple words in the document get automatically differentiated, thereby producing correct frequency count. After this, each component of the document vector- that represents a different universal word (*i.e.*, a concept) is assigned the number of links incident on the node, considering the graph to be undirected. When a UW is not present in the UNL graph of the document then 0 is written in its position. The basic assumption behind this approach of counting the links is that *the more number of links to and from a universal word, the more is the importance of the word in the document.*

3 Document Vector Construction Using UNL Relation Labels

The UNL link method does not consider the label of the links in the graph. In the relation label based method, instead of a single dimensional vector we construct a two dimensional matrix *M* of dimension *n* x *n*, where *n* is the total number of UWs in the corpus encompassing all documents The element m_{ij} of the matrix denotes the value of the weight assigned to the label of the link connecting the UWs, UW_i and UW_j or a value of *0* if there is no link between the two UWs. To make the feature vector we add up all the column of the matrix to form a single dimension vector of size equal to the number of distinct Universal words in the whole corpus. The relation weights are found using a machine learning approach.

4 Evaluation

Vectors of documents were created using the term frequency, the UNL link and the UNL relational label methods. Then they were clustered using the Self Organizing Maps [T. Kohonen 1995]. The neurons were labeled using the majority approach, *i.e.*, if most of the documents assigned to a neuron belong to the cluster *C*, then the label of the neuron is designated as *C*. After the self-organization process, the neurons get labeled and we know the classes of the documents. Then comparing the actual classes with the SOM found classes we can obtain the number of documents correctly clustered. The accuracy of clustering is given by,

$$\textbf{\textit{Accuracy}} = \frac{\textbf{\textit{Number of documents correctly clustered}}}{\textbf{\textit{Total number of documents.}}}$$

4.1 Experiments

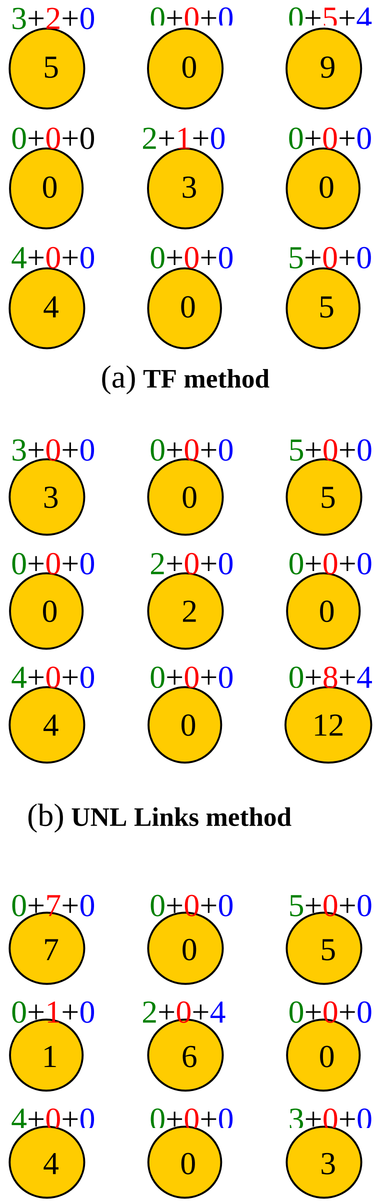
Input:

Total number of documents: 26
Total number of clusters: 3
Documents in cluster 1: 14
Documents in cluster 2: 8
Documents in cluster 3: 4

The Clustering Step:

The dimension of the vector created by TF method for the whole of the twenty-six documents was *1025* and the dimensions of the vectors created by the UNL methods were *1255*. The vectors were then input to a Self Organizing Map of 9 neurons organized as a 3 x 3 grid.

The output of the SOM corresponding to the TF, UNL link and UNL relation method are shown in figures 2(a), 2(b) and 2(c) respectively. The *nine* circles in the figures denote the nine neurons of the 3 x 3 SOM. The number inside the circle denotes the number of documents that were assigned to the neuron *after the self organization process.* The numbers above



(a) **TF method**

(b) **UNL Links method**

(c) **UNL Relation method**

Figure 2: The different Self Organizing Maps

the circles (n_1 + n_2+ n_3) represent the number of documents of class 1, 2 and 3 respectively assigned to that neuron. For example 3+*2+0* above the first circle in figure 1(a) indicates that 3 documents belonging to the first cluster, 2 documents belonging to second cluster and no documents from the third cluster were mapped to that neuron.

5 Discussion of Results

We denote the neurons by the tuple *(row number, column number)* with *row number* increasing from bottom to top and the *column number* increasing from left to right. As seen in figure 2(a), using the term frequency method the documents of clusters 1 are distributed to neurons (1,1), (1,3), (2,2) and (3,1), while those of cluster 2 are given to (3,1), (2,2) and (3,3). The documents of cluster 3 go to (3,3) only. So we have 2+1 documents of cluster 2 and all 4 documents of cluster 3 are wrongly mapped. Hence the accuracy is 19/26 which is 0.730769.

When we consider the UNL link method, figure 2(b) shows that only the 4 documents of cluster 3 are wrongly mapped to the neuron for cluster 2 at (1,3). All 8 documents of cluster 2 are together. The documents of cluster 1- which is big- is distributed to 4 neurons, probably because of intra document differences in spite of being from the same cluster. The accuracy here is seen to be 22/26 which is 0.846154.

Coming to the last method of UNL relation labels, figure 2(c) shows that the distribution of cluster 1 documents are same as before. However, cluster 2 documents stand independently in two neurons. But the good thing is that the cluster 3 now has got an independent neuron label. The number of wrongly clustered documents is only 2 giving, thus, an accuracy of 24/26 which is 0.923077. All the accuracy values are tabulated in table 1.

Method	Accuracy
Term Frequency	*0.730769*
UNL Link	*0.846154*
UNL Relation	*0.923077*

Table 1: Accuracy of different methods

6 Conclusion

We have proposed a new method for text clustering. This method uses the semantic information present in the form of relations between words in sentences. Thus the approach is different from traditional methods of clustering which consider the document as a bag of words. As shown in the experiments, this approach performs better than the methods based on only frequency.

References

A.P. Dempster, N.M. Laird, and D.B. Rubin. *Maximum Likelihood from Incomplete Data via the EM Algorithm*. Journal of the Royal Statistical Society, Series B (Methodological) , 39(1):1--38, 1977.

Julio Gonzalo, Felisa Verdejo, Irina Chugur, Juan Cigarran *Indexing with WordNet synsets can improve Text Retrieval*, Proceedings of the COLING/ACL '98 Workshop on Usage of WordNet for NLP, Montreal. 1998.

J. A. Hartigan and M. A. Wong. *A k-means clustering algorithm.* Applied Statistics, 28:100--108, 1979.

A.K. Jain and R.C. Dubes. *Algorithms for Clustering Data*, Prentice Hall, Englewood Cliffs NJ, U.S.A., 1988.

T. Kohonen *Self-organizing Maps*, Series in Information Sciences, vol. 30, Springer, 1995.

Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, Richard Harshman *Indexing by Latent Semantic Analysis.* Journal of the American Society of Information Science 1990.

Y. Yang, J. O. Pedersen. *A Comparative Study on Feature Selection in Text Categorization*, Proc. of the 14th International Conference on Machine Learning ICML 1997.

H. Uchida, M. Zhu, Senta T. Della. *UNL: A Gift for a Millennium.* The United Nations University, 1995