Emergent Self Organizing Maps for Text Cluster Visualization by Incorporating Ontology Based Descriptors

Kusum Kumari Bharti and Pramod Kumar Singh

Computational Intelligence and Data Mining Research Lab
ABV-Indian Institute of Information Technology and Management Gwalior
Morena Link Road, Gwalior, India
kkusum.bharti@gmail.com, pksingh@iiitm.ac.in

Abstract. Despite various advantages of traditional feature vector model for document representation, the well-known inherent deficiency in this model is “sovereign term assumption”. This deficiency makes it impossible to identify syntactically different but semantically related terms. In this paper, we demonstrate the use of semantic similarity measure for quantifying the relationship between related terms. Identifying such relationships help in reducing the difference between related documents. In this work, we use only noun terms for enriching the representation model. The natural visualization of clusters is investigated in this study using Emergent Self Organizing Map (ESOM). Experimental results show that incorporation of semantic relationship enhances the accuracy of clustering results.

Keywords: Emergent Self-Organized Map (ESOM), feature vector model, semantic similarity measure.

1 Introduction

Text clustering is a more precise subsection of unsupervised learning (also referred as clustering). It is also known as document clustering. It is the process of automatically grouping related documents based upon document’s intrinsic characteristics. It finds its application in real-life applications in various areas, e.g., news aggregation, information retrieval, search engine, topological hierarchy creation.

In current scenario, most documents follow non-linear data structure, i.e., documents deal with large number of diverse areas. For example, “Feature selection algorithm based on particle swarm optimization to improve the performance of document clustering” deals with three different disciplines “Feature selection”, “particle swarm optimization” and “document clustering”. Classical corpus like the well known 20 Newsgroups\(^\text{1}\) and Reuters-21578\(^\text{2}\) dataset deal with a large

\(^{1}\) http://people.csail.mit.edu/jrennie/20Newsgroups/

\(^{2}\) http://www.daviddlewis.com/resources/testcollections/reuters21578/

© Springer-Verlag Berlin Heidelberg 2012
number of diverse topics making the vocabulary very large. In this paper, we use Reuters-21578 dataset for our experimentation.

Vector Space Model (VSM) is a traditional method for representation of documents, where rows correspond to documents and columns corresponds to unique terms/descriptors, or vice versa. Many variations of VSM have been proposed in literature [1]. They differ in what they consider as terms or features. Traditional VSM model uses Bag-of-Words (BOW) representation. The main drawback of this representation is destruction of semantic relations between words. For example, stable phrases such as “Bill Gates” or “White House” are represented in the BOW as separated words so their actual meaning is lost. Given a BOW of a document in which words “White” and “House” occur, one can suggest that the document is somewhat related to color of a house, but not about official residence and principal workplace of the president of the United States. However, given a document representation that contains a phrase “White House”, the reader will hardly make a mistake about the topic of discussion [16].

Traditional VSM model performs strict term based comparison in order to identify related terms and ignores semantic relationship between syntactically different but semantically related terms. For example, “sports” in one document and “athletics” in another document does not contribute to the similarity measure among these two documents. In dealing with inter-disciplinary structure of documents, identifying complex semantic along with reducing dimensions of VSM is striking and open research problem in the area of document clustering.

Self Organizing Map (SOM) [3] is an effective tool for projecting high dimensional non-linear data set onto 1 or 2 dimensional display. Thus, SOM helps in efficiently visualizing inherent non-linear relationships among documents. The most significant benefit of this procedure is that it reduces computational load considerably and also makes it possible to cluster large dataset. These properties make it unique, especially, for dimensionality reduction. These striking features of SOM have provoked researchers to use SOM for text clustering and visualization (e.g., [2]). In ESOM [4] the cluster boundaries are “indistinct”, i.e., the degree of separation between regions of the map (clusters) are depicted by “gradients”. This obviates the need to assign inter-disciplinary documents to more than one cluster. In this paper, we use ESOM for creating clusters of documents and use Wu & palmer [7] semantic similarity measure for identifying complex semantic of terms.

Rest of the paper is organized as follows. Section 2 provides an overview of the relevant literature to understand and present our study. In section 3 we describe our proposed document representation methodology based on frequent legitimated terms and semantically enrich reweighting scheme. Section 4 discusses the experiments and results. Finally, we conclude in section 5 with a hint for possible future research in this direction.

2 Literature Review

There are three main approaches to incorporate terms into the document representation: the first applies unigram only for document representation, the second