**Stats – 517**

**Statistical Learning and Predictive Modeling**

**Literature Review**

**Abhinav Prabhu Adarapuram**

**University of Idaho**

**A Novel Text Clustering Approach Using Deep-Learning Vocabulary Network**

**Abstract**

Mining valuable information on the internet in the organized and collective form is called the Text clustering, the most important to aid the internet surfers. Deep Learning Vocabulary, a novel network approach is constructed eliminating the feature extraction issues and data dimension reductions using related word-set and sparse group deep belief networks. In single pass clustering the coverage rate measure is introduced. The effectiveness of this Novel approach is verified compared to representative algorithms.

**Introduction**

In the present day, we have lots of sources through which many data are created, circulated, collected, shared and deleted. Webpages, blogs, various social medias and networks, electronic gadgets are the major sources for obtaining huge volumes of data. From the available junk of data, text clustering is a valuable text mining method to sort the exact data from the available sources. This clustering methods helps in grouping a huge volume of similar documents or texts together into small and meaningful clusters. This technique is widely used for semantic analysis, text classification, text summarization and topic detections.

The text clustering method is classified into two main categories. They are:

Preprocessing phase

Clustering phase

**Preprocessing Phase:**

Text documents are transformed to structured data in this phase mainly. Text clustering starts with the preprocessing of text and then follows the clustering of text documents. The main objective of the preprocessing phase is to transform the documents into structured data which can then be sent for the clustering process which would be taken care by various algorithms.

Preprocessing is classified into two categories. They are:

Feature extraction

Feature selection

**Clustering Phase:**

The structured data is processed by clustered algorithms including two phases: feature extraction and feature selection.

**Existing Methods**

In concerned with clustering phase where feature extraction is done with the available scientific research literature, two categories are present for feature extraction method.

**1.Term frequency-based method**

This the method of counting words number and term document vectors are extracted from text documents employing vector space method, where each entry of VSM is frequency of word or terms.

However, this method was not effective due to loss of relation of words. Many researchers have found lexical cooccurrence phenomenon, where some words or terms have high probability of occurrence in a text document. The cooccurrence of words or term may result in generating precise feature vectors representing the meaning of text documents.

Feature selection is done to remove redundant information and reduce the dimensionality of term documents. The methods are

1.1 **Corpus-based method**- merges synonyms together reducing dimensionality of features basis large corpora

1.2 **Large Semantic Index** - decomposes term document vector into a term space matrix by singular value decomposition (SVD)

1.3 **Subspace-based** clustering groups text documents in a low dimensional subspace.

**2.** **Semantic Web based method:**

This method is to construct word knowledge in certain domains to an ontology which contains words and their relation.

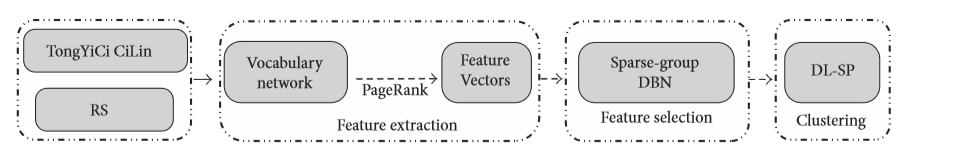
**Idea and Approach**  
  
A novel graph Based Approach- Deep Learning Vocabulary Network is presented for text clustering employing edges of vocabulary network representing the relations between words or terms and extract features of text documents in terms of related word- set, where set is a set of words in the same class and association rules learning to obtain relation between words are used. High dimensional and sparse features of text have big influence on text clustering. Here, deep learning is employed for dimensionality reduction. An improved Deep Learning Single Pass (DL-SP) is used in the process of clustering. The effectiveness of this approach is verified using experimental evaluation based on Chinese corpora.

**Approach of DLVN Proposed:**

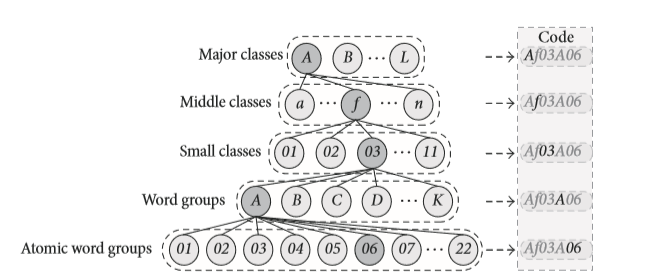
The first step in Deep Learning Vocabulary Network (DLVN) for text clustering is constructing vocabulary network. The nodes of vocabulary network represent words or terms and the edges of vocabulary network to represent relation between words or terms. Two methods in this approach is done to obtain the cooccurrence   relations of words: related word set and TongYoCi CiLin. Frequent item sets are used to discover the relations of items used in database. In DLVN, the related-word sets are created by frequent item sets and each itemset of related-word set is a set of words with cooccurrence relation. Page rank is employed to obtain the "importance" of nodes instead of the term frequency in VSM. An improved DBN (sparse -group DBN) is proposed for dimensionality reduction. For clustering process algorithm, DL-SP is presented in which coverage rate is used for similarity measure.

TongYoCi CiLin is a Chinese semantic dictionary of synonyms and related words, which organizes all words as a five-layer hierarchical tree.

**DLVN Architecture:**



**Structure of TongYoci CiLin:**



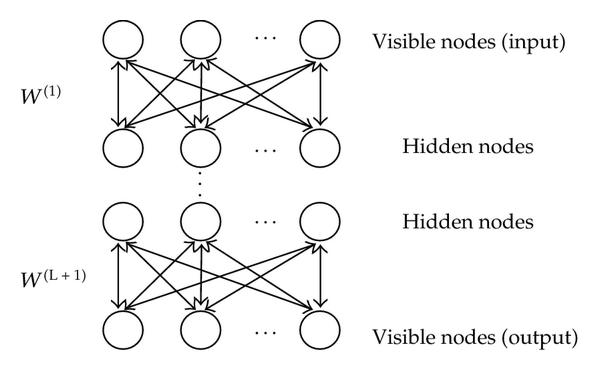
**The Construction of Vocabulary Network:**

It is constructed representing text documents and contains the relation between words or terms. Here, the "importance" of nodes is employed instead of term frequency in VSM.

**Deep-Learning Single-pass:**

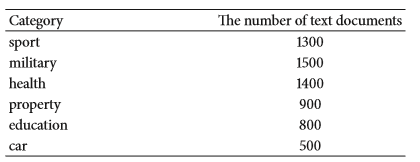
Sparse group DBN is proposed for dimensionality reduction of feature vectors. DBN is a model of Deep Learning. Word dependencies and DBN is combined to propose the sparse group DBN for dimensionality reduction. The coverage rate (CoR) is proposed for similarity measure among feature vectors in DL-SP.

**Structure of DBN**



**Dataset Used**

For analyzing and performing the experiment, we make use of the following list of datasets to identify the effectiveness.



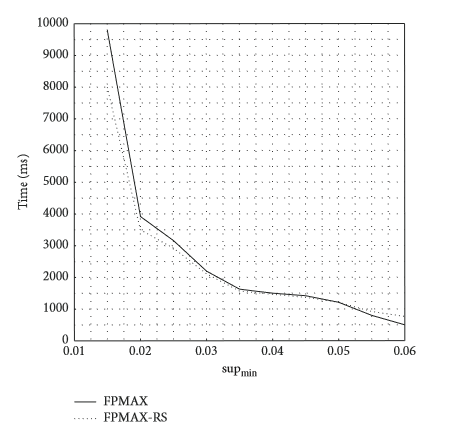
**Results**

The experimental analysis has done on three aspects:

1. The Efficiency of FPMAX-RS in related word set mining is analyzed and found to be resulting in higher efficiency when Sup(min) is smaller

2. The Comparison of Feature vectors is done where the distance among feature vectors is compared based on tf-idc, FC-VSM, and DLVN.

3. Comparison of DL-SP Efficiency, in which DL-SP is compared with LSI and Single Pass which shows DL-SP has better performance compared to LSI and Single Pass.



**Conclusion**

The DLVN approach is proposed for text clustering that involves construction of vocabulary network to mine the importance of words using related word set that contains "cooccurrence" relations of words. The text features of documents in the same category has shorter distance and feature vectors have long distance among different categories. The Sparse group DBN is employed to reduce the dimensionality of feature vectors in terms of the group relations of the words, retaining the word dependency in dimensionality reduction while it must be trained with large group of text documents to improve its performance.

**Future Work**

In current work, we verify the approach using Chinese corpora. In future, English texts will be used to prove the effectiveness.

**References**

A. Trabelsi and O. R. Za¨ıane, “Extraction and clustering of arguing expressions in contentious text,” Data and Knowledge Engineering, *vol.100, pp.226–239,2015.*

K.Schoutenand, F.Frasincar, “Survey on aspect-level sentiment analysis,” IEEE Transactions on Knowledge and Data Engineering*, vol.28,no.3,pp.813–830,2016.*

M.Tsytsarau and T.Palpanas, “Survey on mining subjective data on the web, ”Data Mining and Knowledge Discovery, *vol.24,no. 3,pp.478–514,2012.*

G.Huang, J.He,Y.Zhangetal.,“Mining streams of short text for analysis of world-wide event evolutions,” World Wide Web, *vol.18,no.5,pp.1201–1217,2014.*

U.Erra, S.Senatore, F.Minnella, and G.Caggianese, “Approximate TF-IDF based on topic extraction from massive message stream using the GPU, ”Information Sciences, *vol.292,pp.143– 161,2015.*

C.Qimin, G.Qiao, W.Yongliang,andW.Xianghua,“Text clustering using VSM with feature clusters, ”Neural Computing and Applications, *vol.26,no.4,pp.995–1003,2015.*

J. Martinez-Gil, “An overview of textual semantic similarity measures based on web intelligence,” Artificial Intelligence Review, *vol.42, no.4, pp.935–943,2012.*