**Fuzzy Bag-of-Words Model for**

 **DocumentRepresentation**

**Abstract**

One key issue in text mining and natural language processing (NLP) is how to effectively represent documents using numerical vectors. One classical model is the Bag-of-Words (BoW). In a BoW-based vector representation of a document, each element denotes the normalized number of occurrence of a basis term in the document. To count the number of occurrence of a basis term, BoW conducts exact word matching, which can be regarded as a hard mapping from words to the basis term. BoW representation suffers from its intrinsic extreme sparsity, high dimensionality, and inability to capture high-level semantic meanings behind text data. To address the above issues, we propose a new document representation method named Fuzzy Bag-of-Words (FBoW) in this paper. FBoW adopts a fuzzy mapping based on semantic correlation among words quantified by cosine similarity measures between word embeddings. Since word semantic matching instead of exact word string matching is used, the FBoW could encode more semantics into the numerical representation. In addition, we propose to use word clusters instead of individual words as basis terms and develop Fuzzy Bag-of-WordClusters (FBoWC) models. Three variants under the framework of FBoWC are proposed based on three different similarity measures between word clusters and words, which are named as FBoWCmean, FBoWCmax and FBoWCmin, respectively. Document representations learned by the proposed FBoW and FBoWC are dense and able to encode high-level semantics. The task of document categorization is used to evaluate the performance of learned representation by the proposed FBoW and FBoWC methods. The results on seven real word document classification datasets in comparison with six document representation learning methods have shown that our methods FBoW and FBoWC achieve the highest classification accuracies.

**Architecture**

****

**Existing System**

In topic models including probabilistic latent semantic analysis and latent dirichlet allocation, probability distributions are introduced to describe words and the generation process of each word in a document. The assumption behind topic model is that word choice in a document will be influenced by the topic of the document probabilistically. However, in these models, the derived latent dimension lacks semantic interpretation. For example, LSA regards a latent dimension as a linear combination of all original terms in vocabulary, which is counter intuitive because only a small part of the vocabulary is actually relevant to a certain topic. In addition, these two approaches both utilize the word occurrence of documents to perform dimensionality reduction. However, the occurrence statistics may not be able to capture the true semantic information underlying a document.

**Proposed System**

Different from BoW model and BoW-enhanced models such as LSA and topic models that employ exact word matching and hard mapping, our proposed FBoW and FBoWC models adopt semantic matching and fuzzy mapping to project the words occurred in documents to the basis terms. In our proposed fuzzy BoW models, word embeddings is introduced to help evaluate semantic similarity between words. Since word embeddings are trained on very large-scale corpus, it is believed that the captured similarity information is more accurate and general than that extracted from word occurrence statistics underlying a document in previous BoW-based approaches. In addition, our proposed fuzzy BoW models can also be used in conjunction with the LSA method to reduce the dimensionality of the FBoW representation.

**Future Work**

As a next step work, document structure or word order information will be considered in document representation learning. In addition, the effects of multi-sense word embeddings and different term weighting schemes will be explored in future.

**Module Implementation**

1. **Word Embeddings**

The core idea behind word embeddings is to assign such a dense and low-dimensional vector representation to each word that semantically similar words are close to each other in the vector space. The merit of word embeddings is that the semantic similarity between two words can be conveniently evaluated based on the cosine similarity measure between corresponding vector representations of the two words. In the popular word embeddings word2vec, a twolayer neural network language model is designed to learn vector representations for each word. The word2vec framework contains two separate models including Continuous Bag of Words (CBoW) and Skip-gram with two reverse training goals. CBoW tries to predict a word given the surrounding words, while Skip-gram tries to predict a window of words given a single word.

1. **Fuzzy Bag-of-Words Model**

Firstly, some adopted notations in our proposed methods are introduced. Each document in text corpus is represented by a BoW vector whose elements denote the number of occurrence of basis terms in the document. In a large corpus, only the top l high-frequency words are usually selected as basis terms in BoW model to reduce the sparsity and dimensionality in BoW representations, and the BoW basis terms is therefore a subset of the corpus vocabulary. For traditional BoW representations, documents are mapped into vectors by exact matching of the words in the documents to the basis terms. Exact word matching is equivalent to performing a hard or crisp mapping. If a word w matches a basis term, the output of the crisp mapping function is 1, and is zero otherwise.

1. **Fuzzy Bag-of-WordClusters Model**

It is well acknowledged that BoW model has three limitations, including sparsity, high dimensionality, and lack of capability to encode high-level semantics. The fuzzy BoW model developed in Section III-B addressed the issues of sparsity and semantics, but the high dimensionality problem remains. Actually, the high dimensionality also means redundancy. This is the reason why BoW is often combined with LSA to reduce dimensionality. Certainly, FBoW can also be combined with LSA to reduce the dimensionality and redundancy of FBoW representation. In this study, we propose a plausible method to solve the high dimensionality and redundancy problem of FBoW model. The basic idea of the new method is to use word clusters instead of individual words as basis terms. A word cluster is a group of semantic similar words. Since basis terms are clustered, the dimension and redundancy can be significantly reduced.

1. **Relationships with Previous Text Representation Methods**

Word embeddings are introduced to capture the semantic relationships among words, and the derived semantic similarity and fuzzy mapping are then incorporated into the original BoW model. As a result, the learned document representations are more dense and able to capture more semantic information. In this subsection, we analyze the connections between our proposed FBoW frameworks including FBoW and FBoWC with two typical text representation learning models including dimensionality reduction methods and a deep composition model: convolutional neural network (CNN).

**Algorithm**

1. **K-means clustering algorithm**: Since word embeddings cancapture the semantics of words, K-means clustering algorithmis applied to group these embeddings, i.e., discover wordclusters. The number of clusters, which can be set by users,is equal to the dimensionality of the learned document representation.
2. **Fuzzy Bag-of-Words**: fuzzy Bag-of-Words (FBoW) model isable to reduce sparsity, improve robustness and encodemore semantic information than the original BoWmodel. Instead of binary mapping in BoW, FBoW adoptsfuzzy mapping, in which connections between wordsin documents and basis terms in BoW space are fuzzynumbers. Fuzzy membership value can be determinedaccording to word semantic similarity, which can beeasily calculated based on word embeddings in ourmodel. By introducing fuzzy mapping, our proposedFBoW can be regarded as a more general formulationof BoW.

**SYSTEM REQUIREMENTS**

➢ **H/W System Configuration:-**

➢ Processor - Pentium –IV or Later Version

➢ RAM - 4 GB (min)

➢ Hard Disk - 40 GB

➢ Key Board - Standard Windows Keyboard

➢ Mouse - Two or Three Button Mouse

➢ Monitor - SVGA

**Software Requirements:**

* Operating System - Windows XP or Later Version
* Coding Language - Java/J2EE(JSP,Servlet)
* Front End - J2EE
* Back End - MySQL

**Conclusion**

In this work, we have proposed Fuzzy Bag-of-Words models including FBoW and FBoWC to address issues of sparsity and lack of high-level semantics of BoW representation. Word embeddings are utilized to measure semantic similarity among words and construct fuzzy membership functions of basis terms in BoW space over words in the task-specific corpus. Since word2vec embeddings can be trained over billions of words, word embeddings adopted in our methods are able to capture high-quality and meaningful semantic information that are not contained by the task-specific corpus alone. To determine basis terms in BoW space, FBoWC utilizes word clusters, while FBoW directly regards high term-frequencies words as original BoW does. The adoption of word clusters in FBoWC can reduce feature redundancy and improve feature discrimination. Three different measures have been designed to evaluate similarity between clusters and words, and three corresponding variants of FBoWC models as FBoWCmean, FBoWCmax and FBoWCmin have been developed. The performance of our approaches has been experimentally verified through seven multi-class document categorization tasks.