2008

On a robust document classification approach using TF-IDF scheme with learned, context-sensitive semantics.

Sushain Pandit
Iowa State University, sushain.pandit.isu@gmail.com

On a robust document classification approach using TF-IDF scheme with learned, context-sensitive semantics.

Abstract
Document classification is a well-known task in information retrieval domain and relies upon various indexing schemes to map documents into a form that can be consumed by a classification system. Term Frequency-Inverse Document Frequency (TF-IDF) is one such class of term-weighing functions used extensively for document representation. One of the major drawbacks of this scheme is that it ignores key semantic links between words and/or word meanings and compares documents based solely on the word frequencies. Majority of the current approaches that try to address this issue either rely on alternate representation schemes, or are based upon probabilistic models. We utilize a non-probabilistic approach to build a robust document classification system, which essentially relies upon enriching the classical TF-IDF scheme with context-sensitive semantics using a neural-net based learning component.

Keywords
document classification, neural networks semantics, tf-idf

Disciplines
Artificial Intelligence and Robotics

On a robust document classification approach using TF-IDF scheme with learned, context-sensitive semantics.

Sushain Pandit
Artificial Intelligence Research Laboratory
Department of Computer Science
215 Atanasoff, Iowa State University, Ames, IA 50010
pandit@cs.iastate.edu

Abstract

Document classification is a well-known task in information retrieval domain and relies upon various indexing schemes to map documents into a form that can be consumed by a classification system. Term Frequency-Inverse Document Frequency (TF-IDF) is one such class of term-weighing functions used extensively for document representation. One of the major drawbacks of this scheme is that it ignores key semantic links between words and/or word meanings and compares documents based solely on the word frequencies. Majority of the current approaches that try to address this issue either rely on alternate representation schemes, or are based upon probabilistic models. We utilize a non-probabilistic approach to build a robust document classification system, which essentially relies upon enriching the classical TF-IDF scheme with context-sensitive semantics using a neural-net based learning component.

Introduction

Documents classification comprises automatic assignment of a set of documents into predefined classes by a learning system (classifier) that has been trained on similar data-sets of test documents. Within this context, document indexing is the activity of mapping a document into a form that can be consumed by a classification system. Several document indexing models exist, many of which rely on feature extraction, dimensionality reduction, or both. In feature extraction, the associated document is typically represented as a feature vector encoding presence of words, syntactic entities, or semantically linked tags, and a term-weight is computed for each such feature. This representation, called the bag-of-words approach, or vector space model is decently effective in relatively easier classification tasks and in cases where the documents contains (and are identified by) unambiguous keywords. However, this approach performs poorly when the document contains closely-linked words like synonyms or polysemes. The problem arises from the fact that such schemes ignore key semantic links between words / word meanings and compare documents representations based solely on the word frequencies. Although several approaches have been proposed to address this critical issue over the past decade, even the best classification systems have showed only marginal improvements. In his critical survey on the subject in 2002, Sebastiani hypothesized:

> .. No system is considerably superior to others and improvements are becoming evolutionary. The effectiveness of automated text categorization is unlikely to be improved substantially..

In face of these underlying limitations, one of the possible approaches would be to deviate from inductive learning and achieve a substantially deeper understanding of the document structure using NLP techniques. However, before trying to develop such complete understanding of natural language text, we try to find out the extent of improvement in performance that we can achieve by extending the document representation beyond the contents of the document itself using non-probabilistic approaches. We thus restrict ourselves to the inductive methods and try to enrich the TF-IDF vector with extra words (in vector space) that are in the *context* of the document. These words are acquired by utilizing a neural-net component that has been trained on *(document, word)* pairs to determine whether the word in the pair is in the context of the document. This component is modeled by building upon the notions explained in [1]. In the next section, we discuss some of the related work and prior-art on document classification, followed by our approach. Then, we present some preliminary results and comparisons and finally, conclude the paper with a discussion on research pointers for future work.

Related Work

Within the sphere of general textual classification, various attempts have been made to extend the basic bags-of-words approach. There have been studies to augment it with n-

grams [2] or statistical language models [3]. In the basic vector-space model, a document *T* is represented as a vector instance *V* (or a sparse instance for efficiency) with elements of the form $(g(w_i), i)$, where *g (.)* is a function of the frequency of occurrence of the i-th rooted word (in a dictionary *D)* for the document *T*. One of the common implementation of *g* is given by TF-IDF methodology, which defines $g(w_i)$ w.r.t *T* as:

$g(w_i) = tf_i \cdot \log(N / df_i)$, where tf_i denotes the number of occurrences of w_i in *T*, *N* the total number of documents and df_i denotes the documents in which w_i occurs.

Thus, Tf-IDF essentially stresses the similarity of two documents if they're composed of the same words and as mentioned before, it doesn't take into consideration the semantic links between the words. There have been studies on the use of dimensionality reduction techniques such as Probabilistic Latent Semantic (Probabilistic LSA) Analysis [4], which seeks a k-generative model for word occurrences in a document. This method essentially tries to replace the vector-space model by a latent-space model. Other variants of Probabilistic LSA have been also proposed; however, we won't pursue these further here and would remark the fact that our approach is unique enough in that it focuses on enriching the TF-IDF vector instance with context-specific and semantically relevant words found using a non-probabilistic approach explained in the next section.

Proposed Approach

The approach comprises training of *Neural Network for Text Representation* (NNTR) and *Neural Network for Document Classification* (NNDC), followed by the actual classification task.

Training NNTR Model

The first step is to first train a neural-net on *(document, word)* pairs for a training collection of documents and a dictionary *R* of chosen words. The target for the network is high / low depending on whether *word* is in the *context* of the *document.* This context for the training set is decided by a domain expert, or using one of the approached mentioned in [5]. The simplest way is to define *context* as a containment relation and output high if the word is contained in the document. This network is based on the *Neural Network for Text Representation* model explained in [1]. It uses one-hot encoding scheme for the *word* vectors and TF-IDF for document vectors. Further, it uses three multi-layer perceptrons (MLP), one for word / document vectors each and one to combine the outputs from the first two MLPs. The intent is to relate a distributed and rich representation of words to that of the documents, whenever the word lies within the context of the corresponding document.

Training NNDC Model

Next, we train another neural-net, which would now be used for the actual document classification. We call this *Neural Network for Document Classification*. It is the creation of the input vector (both for training and classification) to this network that makes our system unique when compared with other similar implementations (explained below).

Let *V* denote the vector instance created by applying TF-IDF on an input document *T*. Then, as before:

$$V = \{(g(w_i), i) \mid i \in D\}, \forall \text{ unique } w_i \in T$$

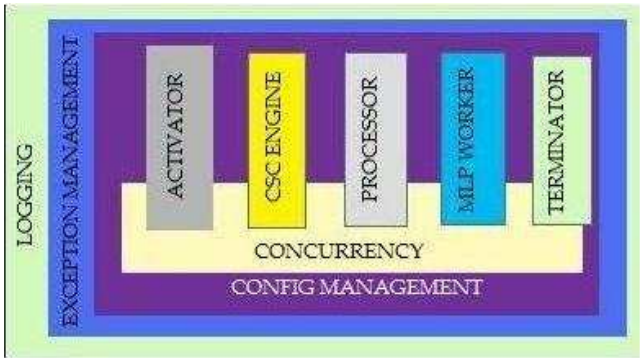
Now, we find out all the semantically *relevant* words in *R* for *T by* invoking $NNTR(w_j, T)$, $\forall w_j \in R$.

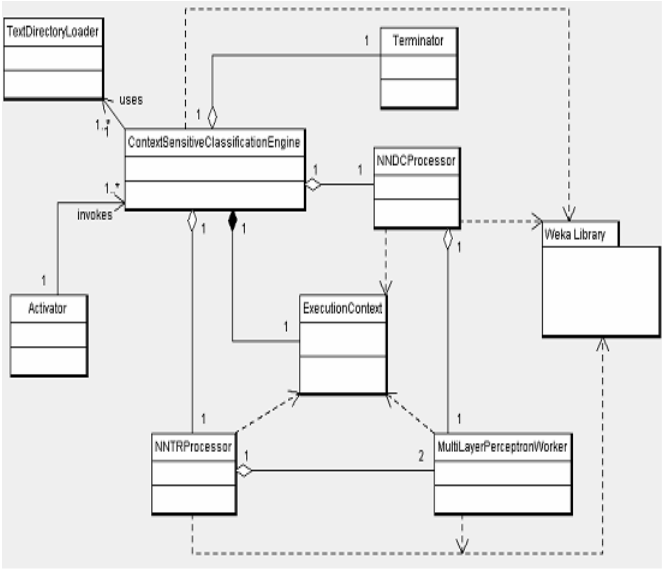
We denote the set of all such *relevant* words by Z and extend *D* by *Z*, i.e., $D = D \cup Z$. Further, we modify *V* to reflect all the words in the set Z, with *g (.)* value equal to a parameter *s* to be adjusted during training phase. Let us denote this modified vector instance as *V'*.

Finally, we train *NNDC* on the modified vector *V'* over all the test documents.

Implementation Details

The system implementation basically utilizes the Weka library API [6]. We use the core functionality from *MultilayerPerceptron* and extend it to create NNTR and NNDC networks and finally hook it all up in *ContextSensitiveClassifier*. We're still working on some of the implementation aspects concerning the TF-IDF filter (WordVector) and parser. A brief architectural overview is provided in the diagrams below.





Preliminary Results

We used the *20NewsGroup* test data samples for out experiments. These data sets contain documents grouped into directories by their class types. We chose *talk.religion.misc* and *alt.atheism* classes with reduced set of 25 documents each, for initial testing. For NNTR training, we manually added relevant keywords related to the two classes. This was relatively hard since the context of the chosen classes is somewhat overlapping (*religion & atheism*). We compared the averaged precision and recall for NNDC for inputs as vector instance *V (simple TF-IDF)* and *V' (TF-IDF* with NNTR). For testing, NNDC had 4 hidden layers, and a dictionary size (*D*) of ~2300. The table below shows the results obtained. We're trying to come up with more comprehensive tests on larger and varied data sets.

Measure	Result on instance vector *V*	Result on instance vector *V'*
Precision	0.816	0.889
Recall	0.708	0.827

Discussion and Future Work

A method for document classification based on an enriched version of TF-IDF model was suggested in this paper. The key idea used was to add some context-specific words to the existing document representation during the training and classification phases. These context-specific words were derived from NNTR model explained in [1]. The overall model is amenable to many enhancements and investigations. It would be interesting to see if by training NNTR on larger clusters of *words* along with *documents* or *words* with subsets of documents (sentences for example), would it be possible to extract more elaborate representations from the documents and train more effectively to classify according to the context. It also needs to be seen how much would this approach scale with larger and more complicated (fuzzier) data sets.

References

[1] Mikaela Keller and Samy Bengio. *A Neural Network for Text Representation.* Artificial Neural Networks: Formal Models and Their Applications - ICANN 2005

[2] Caropreso, Maria Fernanda, Stan Matwin, and Fabrizio Sebastiani. 2001. *A learner-independent evaluation of the usefulness of statistical phrases for automated text categorization*. In Amita G. Chin, editor, Text Databases and Document Management: Theory and Practice. Idea Group Publishing, Hershey, US.

[3] Peng, Fuchun, Dale Schuurmans, and Shaojun Wang 2004. *Augmenting naive Bayes classifiers with statistical language models.* Information Retrieval, 7.

[4] Hofmann, T.: Unsupervised learning by Probabilistic *Latent Semantic Analysis. Machine Learning* 42 (2001)

[5] Evgeniy Gabrilovich and Shaul Markovitch, *Feature Generation for Text Categorization Using World Knowledge.*

[6] Weka 3: Data Mining Software in Java *http://www.cs.waikato.ac.nz/ml/weka/*

[7] Jucheng yang, Chonbuk National University *Implementation of Information Retrieval System with Binary Tree.*

[8] Peter Scheir, Stefanie N. Lindstaedt Know-Center Inffeldgasse 21a, 8010 Graz, Austria. *A network model approach to document retrieval taking into account domain knowledge*

[9] Jae-Ho Kim, Jin-Xia Huang, Ha-Yong Jung, Key-Sun Choi Korea Advanced Institute of Science and Technology (KAIST) / National Language Resource Research Center (BOLA), *Patent Document Retrieval and Classification at KAIST.*