Towards Benchmarking Text Categorization Techniques Using Vector Space Modelling

O. Olaseni
Department of Computer Science
Rufus Giwa Polytechnic
Owo, Ondo State, Nigeria
talktopastor@yahoo.com

ABSTRACT
Text Categorization (TC) is one of the important tasks in information Retrieval (IR) and data mining. The problem of text categorization has therefore become a very active area of research lately as a result of large amount of documents now available on the World Wide Web, in emails and in digital Library. The objective of text categorization is to associate documents with corresponding categories given a large quantity of document in a dataset. It involves building a model from classified documents, in order to classify previously unseen documents as accurate as possible. The intention of our research is to investigate variations of vector space model using inverse document frequently (IDF) and weighted inverse document frequency (WIDF), Using vector space modeling. Our experiment will engage some data sets to be benchmark available approaches to text categorization in order to evaluate the efficiency of the different approaches. The approaches to be benchmarked include the Cosine Coefficient Technique, the Jacquard algorithm, the Dice Coefficient system, the F1 measure and the cosine-Based IDF systems. The findings are expected to contribute to knowledge in the realm of data mining and text categorization techniques.

Keywords: Data Mining, Text Categorizations, Weighted Value, Vector Space and Spam.

1. INTRODUCTION
Text categorization (TC) is one of the important tasks in information retrieval (IR) and data mining. The problem of TC has been active for four decades, and recently attracted many researchers due to the large amount of documents available on the World Wide Web, in emails and in digital libraries. TC involves assigning text documents in a text data collection to one or more of the pre-defined classes/categories based on their content. Unlike manual classifications, which consumes time and require high accuracy, TC makes the classification process fast and more efficient since it automatically categorizes documents.[1]

Many TC strategies from data mining and machine learning (ML) exist such as: decision trees, Support Vector Machine (SVM), rule induction and Neural Network. Our efforts will focus on a text similarity measures to compare the between incoming text (new text cases) and the pre-categorized text in the training data set. Generally, TC based on text similarity goes through two steps: Similarity measurement and classification assignment.

1.1 Term Weighing
Term weighting is one of the known concepts in TC, which can be defined as a factor given to a term in order to reflect the importance of that term. There are many weighting approaches, including, Inverse Document Frequency (IDF) and Weighted Inverse Document Frequency (WIDF). IDF and WIDF focus on terms occurrences inside a text corpus. WIDF distinguishes between two terms that have different occurrences, whereas, IDF treats both terms equally. [1][2]

There are many approaches to assign category to incoming text such as Multiple Associative Classification Methods, Automatic Indexing Based on Bayesian Inference Networks. We intend to implement text-to-text comparison (TTC), which is also known as the K-nearest neighbor (KNN). KNN is a statistical classification approach, which has been intensively studies in pattern recognition over four decades. KNN has been successfully applied to TC problem, and showed promising results if compared with other statistical approaches such as Bayesian based Network. The intention of our research is to investigate variations of vector space model using inverse document frequently (IDF) and weighted inverse document frequency (WIDF).[3]
2. PROPOSED RESEARCH OBJECTIVES

Our research will compare different variations of KNN classification scheme such as Dice, Jacquard and Cosine using IDF and WIDF. The base of our comparison between the different implementations of KNN is the F1 measure. In other words, we want to determine the best SVM, which if merged with KNN produces good result with reference to F1 measures.

3. METHODOLOGY

Using vector space modeling, our experimentation will engage some data sets to benchmark available approaches to text categorization in order to evaluate the efficiency of the different approaches. The approaches to be benchmarked include the cosine coefficient technique, the jacquard algorithm, the Dice Coefficient system, the F1 measure and the Cosine-based IDF systems.

We used F1 evaluation measure as the basis of our comparison, where F1 is computed based on the following equation:

\[ F1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Recall} + \text{Precision}} \]

Precision and recall are widely used evaluation measures in IR and ML, where:

\[ \text{Precision} = \frac{X}{(X+Y)} \]
\[ \text{Recall} = \frac{X}{(X+Z)} \]

The KNN algorithm is a quite simple: Given training and test documents, the algorithm finds the K-nearest neighbors among the training documents, and uses the categories of the k-neighbors to weight the category of the test document. The similarity scores of each neighbor document to the test document are used as a weight of the categories of the neighbor document. If several of the k-nearest-neighbors share a category, then the pre-neighbor weights of that category with respect to the test document. By scoring the candidates’ categories, a ranked list is obtained for the test document. [4]. The various techniques will be compare using two term weighting methods, that is IDF and WIDF in consonance with VSM-based KNN. The Enron Corpus and other dataset will be categorized.

4. EXPECTED CONTRIBUTION TO KNOWLEDGE

The problem of text categorization has a very active area of research lately as a result of the large amount of documents now available on the World Wide Web, in emails and in digital Libraries. Findings from this research will assist in the following:

1) Efficient text categorization that associates documents with corresponding categories given a large quantity of documents in data set.
2) Building a model from classified documents, in order to classify previously unseen documents as accurately as possible.
3) Benchmarking the KNN algorithms such as cosine coefficient technique, the jacquard algorithm, the Dice Coefficient system will equally enable us to know which algorithm is better and in what instances.
4) Finally, our findings are expected to contribute to knowledge in the realm of data mining and text categorization techniques.

REFERENCES