Chapter 60
Local and Global Latent Semantic Analysis for Text Categorization

Khadoudja Ghanem
Constantine 2 University, Algeria

ABSTRACT

In this paper the authors propose a semantic approach to document categorization. The idea is to create for each category a semantic index (representative term vector) by performing a local Latent Semantic Analysis (LSA) followed by a clustering process. A second use of LSA (Global LSA) is adopted on a term-Class matrix in order to retrieve the class which is the most similar to the query (document to classify) in the same way where the LSA is used to retrieve documents which are the most similar to a query in Information Retrieval. The proposed system is evaluated on a popular dataset which is 20 Newsgroup corpus. Obtained results show the effectiveness of the method compared with those obtained with the classic KNN and SVM classifiers as well as with methods presented in the literature. Experimental results show that the new method has high precision and recall rates and classification accuracy is significantly improved.

1. INTRODUCTION

Classification of textual documents (Text categorization) denotes assigning an unknown document (d_i) to a predefined class (C_j / *j=1..N and N is the number of all corpus classes*). Due to the abundance of textual data and with the continuously growing volume of information on the web, categorizing documents into predefined classes makes retrieval, organization, visualization, development and knowledge transfer efficient. Consequently there is an increased interest in developing technologies for automatic text categorization.

Text categorization is a task of text mining (TM), also, it is one of important topics in the fields of Natural Language Processing (NLP), machine learning (ML), information retrieval (IR) and knowledge management (KM), consequently, all used techniques in these different fields can be used to categorize a text.

DOI: 10.4018/978-1-5225-5191-1.ch060

There are several attempts to apply (ML) techniques which include regression models (Yang & Chute, 1994; Schutze et al, 1995), neural networks (Ruiz & Srivinasan, 1998; Acid et al, 2005), probabilistic Bayesian models (Lewis, 1998 & Schneider, 2005), nearest neighbor classifiers, decision trees(Yang & Wang, 2010), symbolic rule learning (Tatiane et al, 2011) and support vector machines(Gunn, 1998; Isa et al, 2008). A detailed survey concerning precedent works can be found in (Fabrizio, 2002; Aggrawal & Zhai, 2012).

In this work, we propose a new text classification approach based on local and global Latent Semantic analysis. The rest of this paper is organized as follows. In section 2 we give an overview on more recent methods presented in the literature which are related to our approach. In Section 3, we present the Latent semantic analysis principle and in section 4 we give the principle of our proposed approach (Classification by Latent Semantic Analysis CLSA). Section 5 discusses obtained results before concluding the paper in Section 6.

2. RELATED WORK

The two main stages in automated document categorization are term reduction and classification. Term reduction is carried by performing feature extraction followed by feature selection. The feature selection methods select a subset of the original set of features (features that have the highest scores) using a global ranking metric (Chi-Squared and Information Gain, for example) or a function of the classifier performance that use a selected feature set. Most authors concentrate their researches on this step, different methods were proposed to reduce terms.

In (Jiang et al, 2012), authors propose an improved KNN algorithm for term reduction, which builds the classification model by combining constrained one pass clustering algorithm and KNN text categorization.

In (Roberto et al, 2012), authors propose a filtering method for feature selection called ALOFT (At Least One FeaTure). The proposed method focuses on specific characteristics of text categorization domain. Also, it ensures that every document in the training set is represented by at least one feature and the number of selected features is determined in a data-driven way.

In (Karabulut, 2013), a two-stage term reduction strategy based on Information Gain (IG) theory and Geometric Particle Swarm Optimization (GPSO) search is proposed with a fuzzy unordered rule induction algorithm (FURIA) to categorize multi-label texts.

A projected-prototype based classifier is proposed in (zhang et al, 2013) for text categorization, in which a document category is represented by a set of prototypes, each assembling a representative for the documents in a subclass and its corresponding term subspace.

Based on k-means clustering feature selection, authors in (Zhou et al, 2014), discuss three classifiers, nearest centroid (NC) method, k nearest neighbor (k-NN) method and SVM method for text categorization. All these methods and many others used different term reduction techniques which lake of semantic. Many other works used semantic techniques in the same step. In (yang & li, 2005), authors developed a system for filtering and recognizing unwanted and harmful messages from legitimate E-mail based on Support Vector Machine (SVM) classification method and on a local LSI representation for each category in order to improve the power of discrimination of categories. With LSI adding to SVM text classification, the classification precision is over about 10-20% than individual SVM.