Chapter 22
Chinese Text Categorization via Bottom-Up Weighted Word Clustering

Yu-Chieh Wu
Ming-Chuan University, Taiwan

ABSTRACT

Most of the researches on text categorization are focus on using bag of words. Some researches provided other methods for classification such as term phrase, Latent Semantic Indexing, and term clustering. Term clustering is an effective way for classification, and had been proved as a good method for decreasing the dimensions in term vectors. The authors used hierarchical term clustering and aggregating similar terms. In order to enhance the performance, they present a modify indexing with terms in cluster. Their test collection extracted from Chinese NETNEWS, and used the Centroid-Based classifier to deal with the problems of categorization. The results had shown that term clustering is not only reducing the dimensions but also outperform than bag of words. Thus, term clustering can be applied to text classification by using any large corpus, its objective is to save times and increase the efficiency and effectiveness. In addition to performance, these clusters can be considered as conceptual knowledge base, and kept related terms of real world.

INTRODUCTION

With the rapid growth of the Internet, there is an increasing need in information technology. Under this situation, creating large corpus becomes much easier than isolated works. In order to effectively deal with the news, organizing text documents is required. The automatic text categorization (TC) is the task of learning to recognize the class label given the testing document. Usually, a machine learning-based classifier is employed to predict the class label. It learns rules from the labeled training data and applied these rules to label testing document. Before applying machine learning algorithms, the document is firstly represented as vectors. In general, the vector is derived from a set of selected words, the so-called bag-of-word (BOW; Uma, Sankar, & Aghila, 2008; Yen, Lee, Wu, Ying, & Tseng, 2011; Galar, Fernan-

DOI: 10.4018/978-1-5225-1759-7.ch022

dez, Barrenechea, Bustince, & Herrera, 2010; Katakis, Tsoumakas, & Vlahavas, 2010; Han & Karypis, 2000). However, the biggest challenge of this approach is that the unknown word is missing and the curse of high dimension. To solve this, a set of dimension reduction-based approaches (Sebastiani, 2002; Bekkerman, El-Yaniv, Tishby, & Winter, 2003; Pereira, Tishby, & Lee, 1993) were proposed over the past years. Examples include, latent semantic indexing (Sebastiani, 2002; Baker & McCallum, 1998), information theoretic clustering (Bekkerman, El-Yaniv, Winter, & Tishby, 2001; Pereira, Tishby, & Lee, 1993). Dhillon et al. (Dhillon, Mallela, & Kumar, 2003) also shows better results with term cluster representation.

Text categorization is a rich and wide research issue. It provides the fundamental step for text mining (Janev, Dudukovic, & Vraneš, 2009; Krogstie, Veres, & Sindre, 2007; Lee, Wu, & Yang, 2009; Nour & Mouakket, 2011; Wu & Chang, Efficient Text Chunking using Linear Kernel with Mask Method, 2007; Wu, Lee, & Yang, Robust and efficient multiclass SVM models for phrase pattern recognition, 2008; Yang, Huang, Tsai, Chung, & Wu, 2009). Some well-known machine learning methods, such as support vector machines (SVM) (Joachims, 1997), had received a great successful in this field. Several well-known machine learning algorithms had been widely used in recent years, for example, SVM (Joachims, 1997), linear classifiers (Widrow-Hoff weighting) (Lewis, Schapire, Callan, & Papka, 1996), Centroid-based learners (Han & Karypis, 2000), memory-based learning (*k* nearest neighbors *k*NN) (Uma, Sankar, & Aghila, 2008; Yang Y., 1999), generalized instant set Widrow (GIS-W) (Lam & Ho, 1998), decision tree (Apte, Damerau, & Weiss, 1994; Sebastiani, 2002); Naïve Bayes (Apte, Damerau, & Weiss, 1994; Bekkerman, El-Yaniv, Tishby, & Winter, 2003), Neural networks (Apte, Damerau, & Weiss, 1994; Uma, Sankar, & Aghila, 2008; Sebastiani, 2002; Yang Y., 1999).

Yang (1999; Yang Y., 1999) gave an earlier work on the comparison of different text classification algorithms. They pointed out that when the training data is sufficient for each category, the four methods, SVM, *k*NN, Neural Nets, and Naïve Bayes yielded no statistical significance. When the data is bias, both SVM and *k*NN showed significantly better accuracy than the other two approaches. Later, Sebastian (2002; Sebastiani, 2002) presented a more detail comparative study and also demonstrated the consistent aspects as well as Yang's work.

Among them, SVM is the most promising algorithm which usually reaches better accuracy than the other methods. However, the main limitation is that it requires great training time to find the optimized instance weights to form the decision hyper-plane. Sometimes, the use of term reduction might not bring better performance (Galar, Fernandez, Barrenechea, Bustince, & Herrera, 2010). Dhillon et al. (Dhillon, Mallela, & Kumar, 2003) indicated that there is no significant change in accuracy when the term clustering technique is applied. Lam and Ho (Lam & Ho, 1998) proposed two novel classification algorithms- GIS-W and GIS-R based on the linear classifiers. They showed very competitive results as well as SVM. It has the advantage in that it is linear and efficient. In testing, the label of the testing document is mainly determined by a set of trained linear classifiers. The *k*NN (Uma, Sankar, & Aghila, 2008; Yang Y., 1999) is one of the famous classification algorithms in the TC field. The main benefit is that it does not need to train. In testing, *k*NN has to compare the testing data with all of the training instances and makes the decision by considering the top-*k* neighbors. As reported by (Yang Y., 1999), the centroid-based algorithm is no significant difference between C4.5 and NB. He also demonstrated that the centroid-based algorithms achieved the best performance in >20 datasets.

To remedy this, we address the curse of high-dimentionality and unknown words for the text categorization. We designed an efficient word clustering algorithm based on the estimation of category distribution between words. Based on the KL-divergence information, it incrementally groups and updates newly