様式８

論　文　内　容　要　旨

報告番号	甲　先　第　280　号	氏　名	Chao Li
学位論文題目	Text Classification Based on Background Knowledge 背景知識に基づく文書分類		

内容要旨

Nowadays, information technology is developing fast and used in more and more field. Therefore, numerous documents are saved in the computers which could be read by the computers. Moreover the number of the documents is increasing extremely everyday. For the important application, it becomes a research subject how to automatically classify, organize and manage such numerous amount of literature and data, in the case most of them are documents.

Predicting class of the online texts has been required by a variety of applications. For example, in spam filtering, classification methods are used to determine the junk information automatically. In news organization, because most news is provide on Internet and the amount is huge, it is impractical to finish this task manually. In emotion classification, a text is classified by the emotion based on their meaning. In recommendation systems, the class of the text would be an important feature determining if its content might draw more attention to this customer.

Depending on the classification task, there are different kinds of class sets. For example, in spam filtering, there are 2 classes in this task, which is a binary classification problem. It only determines if the text is spam mail or not. In news organization, on CNN.com, the news channels include money, entertainment, tech, sport, travel, etc. In emotion classification, there are Joy, Love, Expectation, Surprise, Anxiety, Sorrow, Anger, and Hate.

Due to the growing availability of digital textual documents, automatic text classification (ATC) has been actively studied to organize a vast amount of unstructured documents into a set of categories, based on the textual contents of the document.

Text representation is a fundamental step in text classification task, in which a text could be represented by a set of features. Features play important roles in building classification model and prediction. Most of the previous studies focused on enriching text representation to address text classification task. However, conventional classification approaches with VSM (Vector Space Model) on Chinese text, which only study intensively on the words and their relationship in some specific corpus/dataset.

According to the previous researches, the essential problem of text classification is lack of information, especially for the imbalanced dataset. For example, there two sentences:

Training text: He likes playing basketball.

Test text: I love this game.

There are not any same words in these sentences. We can determine the 1st sentence is related to sport easily. However, if we don't have any background knowledge, we can not know the 2nd sentence is also related to basketball. Because it is a very famous slogan for NBA (National Basketball Association). This shows the effect of background knowledge on text classification.

In this thesis, we illustrate the idea of the background knowledge, which could complement information for documents and build models for text classification task. The motivation for exploiting background knowledge text classification is attributed to two reasons. First, more information from texts can make more reasonable classification. Second, people have basic concept and general knowledge in their mind, however, the conventional corpora/datasets are some kinds of special case which would lack some basic concept and general knowledge. These basic concept and general knowledge is the background knowledge in our life.

This study is based on Baidu Baike and character co-occurrence. Baidu Baike is an online Chinese encyclopedia similar to Wikipedia which is widely used by Chinese speakers to learn basic concept and general knowledge. Two external corpora are employed for extracting the features of character co-occurrence, People's Daily and Sougou news corpus, which are unlabeled corpora.

To predict the categories for new texts, SVM (Support Vector Machine), a machine learning algorithm, is used to build the classification models. The performance of proposed approach is measured on Fudan University text classification corpus and Sougou corpus in reduced version.

The results show that the background knowledge could complement the information for the documents in text classification task. When the corpus is unbalanced, the improvement is obvious with the background knowledge.