Text Classification using hierarchical Structur
Dissertation, 2014

Akmal Saeed Khattak

Abstract

Text Classification is the process of assigning a text document to one or more predefined categories. Patent Classification is one of the application areas of text classification. Patents represent text documents containing detail of industrial technological inventions. Patents are maintained by patent offices. Patents are classified in a hierarchy of categories. International Patent Classification (IPC) is a standard hierarchy maintained by World Intellectual Property Organization) and is used to assign labels to patents. IPC is a complex hierarchical system consisting of various levels and consists of about 80,000 categories distributed across different levels of hierarchy. Manual classification of patents can be quite expensive and laborious given the large number of documents, categories, dimension of text involved and the rules made as a result of manual classification is inconsistent. There is a need to automate the process of patent classification to assign classes to patents in a hierarchy of classes. Some of the challenges in classifying patents automatically are the huge dimension, huge vocabulary, and large number of large documents, large number of categories and many more. This thesis investigates and addresses challenges in automated patent classification. The main focus of thesis is to investigate how better patents can be represented to improve classification accuracy at different levels of IPC hierarchy. For this purpose, different datasets are made based on different fields of patents and order of terms in patents, to investigate its role in improving classification accuracy. One of the focuses of this thesis was on low frequent terms which are considered as noise and are considered to be the worst features for classification in general whereas in patents it can be a technical term and can be a good feature which can better discriminate a class of documents from others. Further, low frequent terms are exploited to make features based on multi-terms that co-occur. In other words, cooccurrence of low frequent terms is exploited. The performance in terms of accuracy in this case is quite significant when compared to single term features. Detailed experiments are performed to investigate what features based on different fields of patents can optimize classification accuracy at different levels of IPC hierarchy. It is also investigated by making a comparison on two datasets in terms of classification accuracy, one on patent dataset and the other on reuters corpus volume 1, to establish the significance of low frequent terms in case of patent dataset. This thesis also investigates the role of clustering in patent classification showing some improvements in performance with respect to classification accuracy.