ABSTRACT

TURKISH DOCUMENT CLASSIFICATION

Rumeysa YILMAZ

M.Sc. Thesis, Department of Mathematics

Supervisor: Assist. Prof. Dr. Rıfat AŞLIYAN

2013, 75 pages

Advancing the technologies of Internet has caused a great deal of digital information and operations. But, because of great amount of digital information, there are some difficulties to reach the information which is stored in databases and processed by systems. In this way, the users in information technology require to reach the information faster and more robust. For that reason, in the classification of the documents in digital technology, the studies about development of new approaches are ongoing. In this study, we aimed to the classification of the documents which are obtained from Turkish web sites.

The documents are categorized to some different classes according to word, stem, syllable and character based approaches. At the same time, n-grams of above units are also used for the classification. The developed systems classify the web based documents with the methods: K-Nearest Neighbor (K-NN), Multi-Layer Perceptron (MLP) and Support Vector Machine (SVM). The systems which use MLP and SVM have been trained and tested. For training and testing operations, two corpora are generated from Turkish web pages. The documents are classified into 6 different classes as "education", "economy", "art and culture", "automobile", "health" and "sport". Each class includes 25 documents and 50 documents in training and testing sets respectively. Thus, the corpora have totally 450 documents for training and testing operations. In preprocessing stage of the system, all unnecessary characters such as punctuation marks are removed from the documents. All capital letters are converted to lower case and only one space character is allowed between two consecutive words. After the frequencies of word, stem, syllable and character n-grams in all documents have been computed in feature extraction stage, the documents have been represented as column vectors which contain frequencies. The tokens as word, stem, syllable and character n-grams are determined with threshold values as 0.25, 0.50, 0.75 and 0.90 in feature selection. In training stage, every document as a feature vector is given to MLP and SVM methods and using these

methods a model is constructed for each class. Finally, the documents in test set are categorized to the classes using the models. The designed systems are evaluated according to the Precision, Recall, Accuracy and F-measure.

The most successful method is SVM with word 1-gram and Accuracy and F-measure score of the systems are 99.9 % and 99.7 % respectively.

Key words: Document Classification, K-Nearest Neighbor, Multi-Layer Perceptron, Support Vector Machine, n-gram.