Categorization of computer science papers using random forest

Zin Mar Oo
University of Computer Studies, Yangon
zmosmilegirl@gmail.com

Abstract

Automatic categorization of text document is difficult and time consuming. This thesis proposed a method for automatic categorization of computer science paper using random forest classifier. Decision trees are widely used for the text categorization task and random forest offers the high accuracy and due to the nature of random forest, it is suitable for the task of text categorization.Random forest used different training dataset and random split at each node for the construction of decision forest.Using random forest can provide high accuracy in categorization of text documents.The system will be implemented using C# programming language MS SQL server 2005. Computer sciences paper from IEEE conferences will be collected and trained using the random forest. Trained random forest will be stored using C# serialization technique in hard disk. Stored random forest can be used to classify the incoming conference paper into their associated category. The proposed system will compute the accuracy on the test dataset using hold-out method and will compare the accuracy with decision tree (C4.5) algorithm.