Abstract

Folksonomy is a non-hierarchical document categorizing system, that treats every category in a flat manner, dan every category is entered freely by anyone who submitted a document in these categories. Categorization is done automatically at the time a document is submitted, by entering the list of categories that best fit the document. del.icio.us (http://del.icio.us) site is one of the most popular social bookmarking sites that uses folksonomy.

Usage of folksonomy, although very easy, also has its weaknesses, such as use of different tags for the same concept, use of the same tag for different concepts, no quality control, etc. We try to provide a solution for some of these problems by analyzing Web documents' contents and categorizing them automatically using multinomial naive Bayes algorithm.

Bayes classifier works by using a set of evidences and a set of classes. By training the system using sample data, we can determine the probability of an evidence given a particular class. Bayes classifier also uses prior probability of a class, which can be calculated from sample data. From these analysis, when given a new document which is formed by a set of evidences (words), the probabilities of each class given that document (posterior probabilities) can be determined.

This system is implemented using PHP 5, Apache, and MySQL. The conclusion from building this system is that the Bayes method can be used to automatically categorize documents and also as an assistive tool for manual categorization.

Keywords: naive Bayes, text classification, folksonomy, indexing