University of Macau

ABSTRACT

HIERARCHICAL CLASSIFICATION OF WEB PAGES

by Wang Yi

Thesis Supervisor: Associate Professor, Zhiguo Gong

Master Degree of Software Engineering

In this thesis, a novel method for web page hierarchical classification is addressed. SVM is used as the basic algorithm to separate any two sub-categories under the same parent node in hierarchy. This hierarchical classification algorithm starts its work from the top of the hierarchical tree downward recursively until it triggers a stop condition or reaches the leaf nodes.

Imbalanced data is a serious problem in real text classification. In order to alleviate the ill shift of SVM classifier caused by imbalanced training data, we try to combine the original SVM classifier with BEV algorithm to create classifier which is called VOTEM. Then, a web document is assigned to a sub-category based on voting from all category-to-category classifiers.

At the same time, the web is growing at an exponential rate and the updating speed of information is incredible from time to time. Therefore, online learning method such as incremental learning is gradually become instrument in practical applications. From our experiments analysis, traditional incremental learning is not excellent in the iterative process. To overcome the drawback caused by using only support vector to represent the whole dataset, we embed some additional information and propose *m-sv-incremental* algorithm to solve this problem.

At last our experiment reveals that two proposed algorithms both obtain better results.