Abstract

Web spamming tries to deceive search engines to rank some pages higher than they deserve. Many methods have been proposed to combat web spamming and to detect spam pages. One basic method is using classification, i.e., learning a classification model from previously labeled training data and using this model for classifying web pages to spam or non-spam. A drawback of this method is that manually labeling a large number of web pages to generate the training data can be biased, non-accurate, labor intensive and time consuming. In this paper, we are going to propose a new method to resolve this drawback by using semi-supervised learning to automatically label the training data. To do this, we incorporate Expectation-Maximization algorithm that is an efficient and an important algorithm of semi-supervised learning. Experiments are carried out on the real web spam data, which show the new method performs very well in practice.

Index Terms

Computer Science

Information Sciences
Keywords
Adversarial Information Retrieval  Web Search  Web Spam Detection
Semi-supervised Learning
Expectation Maximization Algorithm