计算机工程与应用 2010 46 (1): 125-128 ISSN: 1002-8331 CN: 11-2127/TP

数据库、信号与信息处理

一种基于预分类的高效SVM中文网页分类器

许世明[1，2]，武 波[1]，马 翠[2]，邸 思[2]，徐洪奎[2]，杜如虚[2]
1.西安电子科技大学 计算机学院，西安 710071
2.中国科学院 深圳先进技术研究院，广东 深圳 518067

收稿日期 2008-7-23 修回日期 2008-10-23 网络版发布日期 2010-1-7 接受日期

摘要 中文网页分类技术是数据挖掘研究中的一个热点领域，而支持向量机（SVM）是一种高效的分类识别方法。首先给出了一个基于SVM的中文网页自动分类系统模型，详细介绍了分类过程中涉及的一些关键技术，其中包括网页预处理、特征选择和特征权重计算等。提出了一种利用预置关键词表进行预分类的方法，并详细说明了该方法的原理与实现。实验结果表明，该方法与单独使用SVM分类器相比，不仅大大减少了分类时间，准确率和召回率也明显提高。

关键词 支持向量机 中文网页分类 文本分类 机器学习
分类号 TP391.1

Efficient SVM Chinese Web page classifier based on pre-classification

XU Shi-ming[1，2]，WU Bo[1]，MA Cui[2]，DI Si[2]，XU Hong-kui[2]，DU Ru-xu[2]

1.School of Computer Science and Technology，Xidian University，Xi’an 710071，China
2.Shenzhen Institute of Advanced Technology，Chinese Academy of Sciences，Shenzhen，Guangdong 518067，China

Abstract
Chinese Web page classification has been considered as a hot research area in data mining，and SVM is an effective method for learning the classification knowledge from massive data.In this paper，a model of automatic Chinese Web page classification system based on SVM is presented first.Then detailed design and implementation are introduced，and some key techniques about Chinese Web page classification，including Web page pre-processing，feature selection and weight computing are discussed.A pre-classification method by a given keywords list is proposed，and the principles and detailed implementation are described.The experiment shows that it not only reduces time but also increases in precision and recall compared with using SVM classifier only.

Key words support vector machine Chinese Web page classification text classification machine learning

DOI: 10.3778/j.issn.1002-8331.2010.01.039

通讯作者 许世明 smingxu@qq.com