山东大学学报(工学版) 2013, 43(2) 29-34 DOI: ISSN: 0412-1961 CN: 21-1139

本期目录 | 下期目录 | 过刊浏览 | 高级检索 [打印本页] [关闭]

机器学习与数据挖掘

基于独立成分分析和协同训练的垃圾网页检测

高爽1,2,张化祥1,2*,房晓南1,2

1. 山东师范大学信息科学与工程学院,山东 济南 250014;
2.山东省分布式计算机软件新技术重点实验室, 山东 济南 250014

摘要:

垃圾网页检测具有重要意义,由于只有少量标记网页,所以可使用半监督协同训练方法检测垃圾网页。将网页特征分为两个视图,即内容视图与链接视图。首先使用独立成分分析分别提取两视图特征的独立成分,然后进行协同训练。实验结果表明,该方法可有效提高垃圾网页检测精度,同时验证了对两个视图分别进行独立成分分析相比于其他方法更为有效。

关键词: 多视图分类 独立成分分析 协同训练 垃圾网页检测

Independent component analysis and co-training based Web spam detection

GAO Shuang1,2, ZHANG Hua-xiang1,2*, FANG Xiao-nan1,2

1. Department of Information Science and Engineering, Shandong Normal University, Jinan 250014, China;
2. Shandong Provincial Key Laboratory for Novel Distributed Computer Software Technology, Jinan 250014, China

Abstract:

Web spam detection is of great significance, and there only exists a small number of labeled pages. Thus, the semi-supervised co-training was used to detect the Web spam pages. The page features were divided into two views, the content view and the link view. First, the independent components of each view were extracted by the independent component analysis, and then the co-training was used to detect the label of each Web page. Experimental results showed that this method could effectively improve the recognition accuracy of Web spam. The results also verified that two respective independent component analyses of each view were more effective than the other methods.

Keywords: multi-view classification independent component analysis co-training Web spam detection

收稿日期 2012-12-05 修回日期 网络版发布日期

DOI:

基金项目:

国家自然科学基金资助项目(61170145);教育部高等学校博士点专项基金资助项目(20113704110001);山东省自然科学基金资助项目(ZR2010FM021)

通讯作者: 张化祥(1966-),男,山东济宁人, 教授, 博士生导师, 主要研究方向为机器学习,模式识别及Web挖掘等. E-mail: huaxzhang@163.com

作者简介: 高爽(1988-),女,山东济南人,硕士研究生,主要研究方向为机器学习与模式识别.E-mail: 824223485@163.com
作者Email:

PDF Preview

扩展功能
本文信息
Supporting info
PDF(1728KB)
参考文献[PDF]
参考文献
服务与反馈
把本文推荐给朋友
加入我的书架
加入引用管理器
引用本文
Email Alert
文章反馈
浏览反馈信息
本文关键词相关文章
多视图分类
独立成分分析
协同训练
垃圾网页检测
本文作者相关文章
PubMed

参考文献：