基于异构特征和组合分类器的网页分类

邓立，杜歆，沈继忠
浙江大学信息与电子工程学院，中国杭州市，310027

摘要：网页特征是网页分类的关键，通过有区分度的特征能有效对网页分类。网页结构特征是对文本特征的有效补充。不同分类器有不同特点，多分类器组合可实现分类器性能互补。提出一种基于异构特征和组合分类器的网页分类算法。与计算HTML标记的频率不同，本文采用树状分布的HTML标签表示网页结构特征，以向量形式将异构文本和结构特征融合。通过计算一组样本的分类准确率，提出将分类结果置信度作为比较不同分类器分类结果的标准。基于置信度采用投票、比较大小和直接输出的决策策略，得到组合分类器的分类结果。实验结果表明，在Amazon数据集、7-web-genres数据集和DMOZ数据集中，准确率分别提高到94.2%、95.4%、95.7%。融合文本和结构特征的分类方法比仅使用文本特征的方法更全面有效。同时多分类器组合能够提高网页分类准确率，高于同类网页组合分类算法。

关键词：网页分类；网页特征；分类器组合
https://doi.org/10.1631/FITEE.1900240