计算机应用 2006 26 (12): 3012-3014 ISSN: 1001-9081 CN: 51-1307/TP

软件开发与典型应用

一种基于锚文本和改进C4.5决策树算法的主题爬行方法

刘金红[1]; 陆余良[1,2]

解放军电子工程学院网络系[1]

收稿日期 2006-6-5 修回日期 2006-7-26 网络版发布日期 2006-12-25 接受日期

摘要 提出了一种基于锚文本和改进C4.5决策树算法的主题爬行方法：基于锚文本词项集训练决策树，然后基于决策树模型来计算网页的主题相关性和待爬行URL的优先级顺序。最后，应用该方法在四所大学网站网页数据集上针对“学术报告”主题进行了主题爬行实验，并与两种标准的网络爬虫进行了性能对比，实验结果验证了该方法的有效性。

关键词 主题网络爬虫 锚文本 决策树

分类号

DOI:

对应的英文版文章：6063112

通讯作者:
刘金红 wondergoldff@gmail.com
作者个人主页: 刘金红 陆余良