计算机应用 2008 28 (04): 942-944 ISSN: 1001-9081 CN: 51-1307/TP

人工智能

一种基于HITS的主题敏感爬行方法

蒋宗礼[1];徐学可[1];李帅[2]

北京工业大学计算机学院[1]
清华大学电子工程系[2]

收稿日期 2007-10-25 修回日期 2007-11-10 网络版发布日期 2008-4-28 接受日期

摘要 基于主题的信息采集是信息检索领域内一个新兴且实用的方法,通过将下载页面限定在特定的主题领域，来提高搜索引擎的效率和提供信息的质量。其思想是在爬行过程中按预先定义好的主题有选择地收集相关网页，避免下载主题不相关的网页，其目标是更准确地找到对用户有用的信息。探讨了主题爬虫的一些关键问题,通过改进主题模型、链接分类模型的学习方法及链接分析方法来提高下载网页的主题相关度及质量。在此基础上设计并实现了一个主题爬虫系统，该系统利用主题敏感HITS来计算网页优先级。实验表明效果良好。

关键词 主题爬虫 超链接引导的主题搜索 主题模型

分类号

DOI:

对应的英文版文章:A7105855

通讯作者:
蒋宗礼 jiangzl@bjut.edu.cn
作者个人主页: 蒋宗礼 徐学可 李帅