

开发研究与设计技术

互联网商品信息抽取技术

于鲁波¹, 陈超²

(1. 中国科学技术大学电子工程与信息科学系, 合肥 230027; 2. 多媒体计算与通信教育部微软重点实验室, 合肥 230026)

收稿日期 修回日期 网络版发布日期 2008-3-3 接受日期

摘要 针对网页信息抽取中格式多样化的问题, 提出一种基于路径统计聚类的信息抽取算法。该算法充分利用电子商务网站网页的特点, 给出网页统计信息的一般数学表达式, 在此基础上, 采用基于统计聚类的思想, 分割信息块, 实现抽取信息。通过对实际电子商务网站网页信息的抽取, 证明算法的有效性, 分割正确率达92.27%, 信息抽取正确率达98.24%。

关键词 [网页分割](#) [网页信息抽取](#) [包装器](#) [路径聚类](#)

分类号 [TP391](#)

DOI:

通讯作者:

作者个人主页: 于鲁波¹; 陈超²

扩展功能

本文信息

▶ [Supporting info](#)

▶ [PDF](#) (123KB)

▶ [\[HTML全文\]](#) (0KB)

▶ [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](#)

相关信息

▶ [本刊中 包含“网页分割”的 相关文章](#)

▶ 本文作者相关文章

• [于鲁波¹, 陈超²](#)