

软件技术与数据库

基于数据区域发现的信息抽取规则生成方法

曲著伟^{1,2}, 李敏强¹

(1. 天津大学管理学院, 天津 300072; 2. 浙江财经学院信息学院, 杭州 310018)

收稿日期 修回日期 网络版发布日期 接受日期

摘要 提出一种自动检测网页中数据记录结构特点并生成Web信息抽取规则的方法, 以网页DOM 树为基础, 自动发现和分离Web数据区域所对应的DOM子树, 将其分解为数据记录子树集合, 综合数据记录子树的结构特点生成抽取规则。实验结果显示, 该方法具有较高的抽取准确率和查全率。

关键词 [信息抽取](#); [抽取规则生成](#); [Web数据区域](#); [树匹配](#)

分类号 [TP311.12](#)

DOI:

通讯作者:

作者个人主页: [曲著伟^{1;2};李敏强¹](#)

扩展功能

本文信息

▶ [Supporting info](#)

▶ [PDF\(141KB\)](#)

▶ [\[HTML全文\]\(0KB\)](#)

▶ [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“信息抽取; 抽取规则生成; Web数据区域; 树匹配”的相关文章](#)

▶ [本文作者相关文章](#)