

软件技术与数据库

一种改进的基于本体的Web信息抽取

柳佳刚, 陈山, 黄樱

(湖南工学院计算机科学系, 衡阳 421002)

收稿日期 修回日期 网络版发布日期 接受日期

摘要 以Web页面信息项本体定义为基础, 对单个样本页面信息项路径进行启发式学习, 对所有样本页面集中信息块路径进行归纳学习, 识别结构相似的信息块子树位置, 以准确划定信息抽取区域, 降低页面噪声。将经过噪声处理的样本页面自动解析成页面的结构本体。比较Web页面信息项本体和页面的结构本体, 通过归纳学习算法生成抽取规则, 提高Web信息的抽准率。

关键词 [信息抽取](#); [本体](#); [归纳学习](#); [文档对象模型](#)

分类号 [N945](#)

DOI:

通讯作者:

作者个人主页: [柳佳刚](#); [陈山](#); [黄樱](#)

扩展功能

本文信息

- ▶ [Supporting info](#)
- ▶ [PDF\(101KB\)](#)
- ▶ [\[HTML全文\]\(0KB\)](#)
- ▶ [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“\[信息抽取; 本体; 归纳学习; 文档对象模型\]\(#\)”的 \[相关文章\]\(#\)](#)
- ▶ [本文作者相关文章](#)