计算机工程 2009 35 (22): 59-61 ISSN: 1000-3428 CN: 31-1289/TP 首页 当期目录 上一期

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