

开发研究与设计技术

基于网页框架和规则的网页噪音去除方法

时达明, 林鸿飞, 杨志豪

(大连理工大学计算机科学与工程系, 大连 116024)

收稿日期 修回日期 网络版发布日期 2007-9-28 接受日期

摘要 提出了一种基于网页框架和规则的网页去除噪音的新方法, 该方法根据网页中HTML标签<table>将网页分成若干部分, 对各个table的长宽比属性进行比较, 去掉长宽比很大的部分, 并对其余table中的内容进行分析, 根据内部是否存在和段落文字有关的标签<p>或
等来区分主题内容和噪音内容, 在此基础上去除噪音内容。对来自CWT200G语料的132 559个网页进行测试后的结果表明, 该方法可以有效地去除网页噪音, 使索引文件减少约75%, 大大地提高了检索速度, 准确度也得到一定提高。

关键词 [信息检索](#) [网页噪音](#) [页面框架](#)

分类号 [TP393](#)

DOI:

通讯作者:

作者个人主页: 时达明; 林鸿飞; 杨志豪

扩展功能

本文信息

- ▶ [Supporting info](#)
- ▶ [PDF\(188KB\)](#)
- ▶ [\[HTML全文\]\(0KB\)](#)
- ▶ [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“信息检索”的 相关文章](#)
- ▶ 本文作者相关文章
- [时达明, 林鸿飞, 杨志豪](#)