现代图书情报技术 2007 (1): 40-43 ISSN: 1003-3513 CN: 11-2856/G2

20届机检会论文选登

一种通用HTML网页主题信息提取方法*

许文[1];都云程[1];李渝勤[1];施水才[1,2]

北京信息科技大学中文信息处理研究中心[1]

收稿日期 2006-10-9 修回日期 网络版发布日期 2007-1-24 接受日期

摘要 采用DOM规范，把HTML网页表示成树结构，对不同模板的HTML页面“主题”信息提取进行研究和分析，提出一种新的结点主题相关性判定方法，依据此方法判定出要抽取的主题内容，并删除无关内容，结果输出只含主题信息的HTML文档。

Abstract By researching how to extract the topical contents in different kinds of templates of Web pages, this paper introduces a new extraction methodology based on DOM. The approach transforms HTML documents into DOM trees. According to the method, the topical contents are extracted and topic-unrelated content are deleted. The result of the approach represents the HTML document which only contains the topic information.

关键词 DOM 信息提取 分块 相关度

Key words DOM;Information extraction;Partition;Correlativity

分类号 TP391

DOI:

通讯作者:
许文 xu.wen@trs.com.cn
作者个人主页: 许文 都云程 李渝勤 施水才