计算机工程与应用 2007 43 (36): 169-172 ISSN: 1002-8331 CN: 11-2127/TP

数据库与信息处理

一种互联网新闻网页的采集分析方法

吴定明, 赵东岩

北京大学 计算机研究所，北京 100871

收稿日期 修回日期 网络版发布日期 2007-12-10 接受日期

摘要 设计了一种采集分析互联网新闻网页的方法。该方法根据给定的新闻网站的入口地址在网络上找出所有的相关链接；区分这些链接所指向的页面特征，过滤掉相关性不大的内容，提取所有新闻网页的链接；进而进行多层次链接分析，根据新闻的图片、标题字体属性及日期，采用NewsPageRank算法计算每个新闻链接的权重。测试结果表明该方法对Internet上的新闻站点普遍具有较好的分析效果，性能可以满足实用要求。

关键词 链接分析 页面评估 互联网 网页采集 链接识别 链接权重 网页权重分析 新闻网页

分类号

Method of collecting and analyzing news pages on Internet

WU Ding-ming,ZHAO Dong-yan

Institute of Computer Science & Technology，Peking University，Beijing 100871，China

Abstract

This paper gives a method of collecting web pages of news.That is downloading the entry web page of a specified website，distinguishing the characters of the pages to which the entry web page links，filtrating irrelevant contents and extracting all the correlative hyperlinks of news on the entry web page.Considering the style of titles，the pictures and date of news，the method analyzes multi-levels hyperlinks and gives the ranking of those hyperlinks using NewsPageRank algorithm.The result of testing shows that the method adapts to the majority of websites of news and has a good practicality.

Key words analyze hyperlinks PageRank Internet collect web pages identify hyperlinks hyperlink weight analyze page weight news page

DOI:

通讯作者 吴定明 wudingming@icst.pku.edu.cn