山东大学学报(理学版) 2007, 42(9) 62-66 DOI: ISSN: 1671-9352 CN: 37-1389/N

论文

基于网页分块的Shark-Search算法

陈 军[1]，陈竹敏[2]

1. 山东大学网络中心， 山东 济南 250100； 2. 山东大学计算机科学与技术学院， 山东 济南 250061

摘要：

Shark-Search算法是一个经典的主题爬取算法. 针对该算法在爬取噪音链接较多的Web页面时性能并不理想的问题, 提出了基于网页分块的Shark-Search算法, 该算法从页面、块、链接的多种粒度来更加有效的进行链接的选择与过滤. 实验证明, 改进的Shark-Search算法比传统的Shark-Search算法在查准率和信息量总和上有了质的提高.

关键词： Shark-Search算法 主题爬取 页面分块 相关性计算

Improved Shark-Search algorithm based on page segmentation

CHEN Jun[1]，CHEN Zhu-min[2]

1. Network Center, Shandong University, Jinan 250100, Shandong;2. School of Computer Science and Technology, Shandong University, Jinan 250061, Shandong

Abstract:

A Shark-Search algorithm is one of the classical algorithms for focused crawling. However, its performance is not ideal for crawling Web pages which contain too many noisy links. An improved Shark-Search algorithm based on page segmentation was proposed, which can accurately evaluate the relevance from three granularities: page, block and single link. Several experiments were carried out to verify that the improved Shark-Search algorithm can obtain significantly higher efficiency than traditional ones.

Keywords: Shark-Search algorithm focused crawling page segmentation relevance computation

收稿日期 1900-01-01 修回日期 1900-01-01 网络版发布日期 2006-10-24

DOI:

基金项目:

通讯作者: 陈 军

作者简介:

本刊中的类似文章

1. 苏 祺,项 锟,孙 斌 .基于链接聚类的Shark-Search算法[J]. 山东大学学报(理学版), 2006,41(3): 1-04

扩展功能

本文信息

Supporting info

PDF(362KB)

[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

本文关键词相关文章

- Shark-Search算法
- 主题爬取
- 页面分块
- 相关性计算

本文作者相关文章

- 陈 军
- 陈竹敏