网络引文的爬虫系统初探

摘要

随着互联网技术的高速发展，网络信息资源越来越受到人们的重视，成为人们获取信息和进行科学研究的主要信息来源。网络引文，即网络参考文献也逐渐受到众多学术工作者的青睐，并成为一种不可忽视的引文类型。统计分析网络引文，不仅可以了解各领域对网络引文的利用现状，而且可以对其进行系统分析、数据挖掘。本文以万方数据库为例，通过爬虫拉取部分论文的网络引文，并进行解析、存储，为后续分析数据提供了强有力的数据保障。

关键字：网络引文；引文分析；数据挖掘；网络爬虫

Abstract

目录

1. 绪论
   1. 论文的研究背景及意义
   2. 网络爬虫系统的研究现状
   3. 论文的相关工作
   4. 论文的安排
2. 相关概念及技术背景
   1. 网络引文的概念
   2. 网络爬虫的概念
   3. 网络引文爬虫
      1. 网络引文的爬虫与通用爬虫的区别
      2. 网络爬虫的工作原理
   4. 本章小结
3. 网络引文爬虫系统ReferenceSpider的设计
   1. 网络引文爬虫系统ReferenceSpider的总体设计
   2. URL提取
   3. URL消重
4. 网络引文爬虫系统ReferenceSpider的实现
   1. 系统环境
   2. 测试报告
   3. 实验结论
   4. 本章小结
5. 总结与展望
   1. 本文总结
   2. 下一步工作

正文（前言、本轮、结论）

致谢

参考文献

1. 绪论
   1. 论文的研究背景及意义

近年来，计算机和信息技术有了突飞猛进的发展，产生了许多新概念和技术，如更高性能的计算机和操作系统、因特网（internet）、数据仓库（datawarehouse）、神经网络等等。在市场需求和技术基础都满足的情况下，数据挖掘的概念和技术应运而生。

数据挖掘（Data Mining）旨在从大量的、不完全的、模糊的随机数据中提取出人们事先不知道的潜在的有价值的信息或知识。http://www.xuexila.com/baikezhishi/1244762.html

网络的迅速发展使各行各业都开始采用计算机及相应的信息技术进行管理和运营。网络引文由于其传播方式的自由性、动态性，也逐渐成为论文参考文献的主要来源之一。为了能更方便的对论文中引用到的网络引文进行数据分析和数据挖掘，我们需要更方便的收集论文的网络引文，然而由于数据量较大、手工操作复杂，导致网络引文分析所需的数据供给不足，这时就需要一个网络引文的爬虫系统来拉取数据，提供数据分析和数据挖掘所需的海量数据。

* 1. 网络爬虫系统的研究现状

网络爬虫系统主要是用来在浩瀚网络中收集网页或特定类型的数据信息，然后从中解析出客户所需要的信息，并持久化到数据库系统。

目前在国内外存在很多互联网爬虫系统，很多企业也会利用爬虫系统定时爬取所需数据。

基于web的爬虫系统是最普遍的，从某个URL作为根路径，然后对整个互联网节点进行信息采集，适用于广泛的主题搜索应用。采集数据量大，并且缺少针对性。这类爬虫代表系统有Mercator和Google。

另一种基于Agent的互联网爬虫系统，是利用Agent技术在信息采集的同时，对用户倾向进行观察和记录，然后不断调整数据采集策略。Agent爬虫系统能够智能、灵活地处理某些个性化信息采集。这类爬虫的代表系统有麻省理工学院研究开发的Letizia、Amalthaca、美国爱荷华大学的ARACHNID。

还有一种是基于增量的采集系统，主要用来采集那些数据更新频繁的网页。工作原理是：在完成初次数据采集后，后续的采集只针对那些改变了的数据，对于未改变的数据将不再处理。增量采集方式可以提高数据爬取的效率，最大程度地提高网络和磁盘等资源的利用率。代表系统有：天网增量搜集系统、Univ、WebFountain、Chile Crawler。

网络舆情爬虫系统的设计与实现 李海燕

* 1. 论文的相关工作
  2. 论文的安排

1. 相关概念及技术背景
   1. 网络引文的概念
      1. 引文

引文是现代学术论文的重要组成部分，它表征了本文献与被引用文献之间的引用关系。在学术研究过程中，任何一篇论文的写作，都离不开已有文献或资源的借鉴，这里的借鉴可以理解为我们通常所说的引用。引用的内容称为引文。

引文，又称参考文献，出现的形式有两种，一是作为附属内容置于文章或者论文的后面，二是作为文章内容的注释存在。引文是作者在撰写文章时由于参考并引用了他人的有借鉴意义的观点而形成的著录方式。

引文具有以下功能：

1. 体现作者的写作态度以及反映论文具有真实的科学依据；
2. 表明作者对于前人成果的尊重以及展示自己的创新和发展；
3. 为读者提供所引用参考文献以及作者简介的出处；
4. 表明学科的发展以及与其他学科之间的联系；
5. 为情报研究和文献计量研究提供依据。

因此，引文已经成为学术论文中不可缺少的内容。

网络引文分析和传统引文分析的比较 马晓佳

国外网络引文研究的现状及展望 杨思洛

* + 1. 网络引文

网络引文（也称网络参考文献，Webcitation,Internet-based citation,Web-based citation）就是将网络文献作为学术论文参考文献的一种引文形式。网络文献是一种虚拟信息化资源，以计算机存储介质或网络为载体，传播交流形式自由，大多为超文本链接，具有动态性。

然而，不是所有的网络文献都可以成为网络引文。要成为网络引文，需要必要的学术认可，即要具有信息真实性、严肃性和资源的稳定性。在网络环境中，网络的开放性和自由度导致了其信息的不确定性和不规范性，从而无法保证信息真实性和严肃性。虽然从学术研究的角度看，网络还未形成规范的学术交流体系，但网络的快速发展正猛烈冲击传统的学术交流体系，网络引文的大量使用已经无法避免。因此论文著作者在引用网络文献之前需要对其真实性、严肃性、稳定性进行评估，合理选择引用的网络文献。另外网络文献的开发者和建设者应该努力为网络资源的利用提供高效可行的评价指引机制，使其早日正式化、规范化。

网络引文信息的挖掘及其对数字图书馆资源建设的启示 王建芳

* 1. 网络爬虫技术