江苏师范大学

毕业设计开题报告

论 文 题 目：网上舆情爬取系统的设计与实现

姓　　名：李林

学　　院：智慧教育学院

专　　业：软件工程

年 级、学 号：2012 、12267067

指 导 教 师：董永权

江苏师范大学教务处印制

毕业设计开题报告

研究目的和意义:

随着互联网技术的发展与应用的普及，网络作为信息的载体，已经成为社会大众参与社会生活的一种重要信息渠道。由于互联网是开放的，每个人都可以在网络上发表信息，内容涉及各个方面。小到心情日志，大到国家大事。互联网已成为思想文化信息的集散地，并具有传统媒体无法相比的优势：便捷性、虚拟性、互动性、多元性。

网络舆情热点通常形成迅速，多是人们对于日常生活中的各种问题发表的各种意见，评论，态度，情绪等，随着事件的发展而变化，是反映社会热点的重要载体之一。

网络爬虫是一种按照一定上网规则，自动的抓取万维网信息的程序或脚本。网络检索功能起于互联网内容爆炸性发展所带来的对内容检索的需求。搜素引擎不断发展，人们的需求也不断提高，网络信息搜索已经成为人们每天都有进行的内容。本课题来源于舆情爬虫系统项目的设计与实现，旨在为相关机构提供及时的网络信息服务。这些服务与现有的搜索引擎提供的服务不同，其重要特征主要体现在：及时性，专用性，人性化。

本文完成的是舆情爬虫系统的设计与实现，该爬虫系统为舆情分析系统提供数据源，完成舆情信息的搜集。因此可以说爬虫系统是整个分析系统的基础，并且爬虫系统输出结果的好坏直接影响着系统结果的展现。

课题研究现状：

从搜索对象上来分类，主流的的爬虫技术包括以下两种：

第一种是基于链接分析的搜索。上世纪九十年代，国外的搜索引擎开发者已经开始以社会网络工作为模型，对万维网进行模拟。专家们通过社会间人与人的关系网，设计研发出了页面间的超链接关系网络。同时他们还惊奇的发现，相似度最高的在传统引文方面。这样通过对照就可以分析得出结论，从关系网络的角度入手，就能将互联网上大量的网页进行分类。早在 2002 年，欧美地区便出现了这种最原始的基于链接的搜索系统。

第二种是基于内容分析的搜索。相对于基于链接分析的搜索方式，这是搜索技术的一个突破性进展，他们采取了一种新的思维方式，建立一个针对主题的词库。当用户在专业领域进行搜索时，可以将词库和爬虫结合起来进行检索。由于搜索角度的转变，这种新的技术逐渐开始被人们所关注。

对于基于内容分析的搜索，国人也做出了很大贡献。张福炎教授设计出可以对万维网上的中英文内容进行搜索，大大的填补了中文方面的空白。它能够在万维网上对信息进行自动查询，采用向量空间模型技术对内容进行检索，同时利用权重评价技术来进行统计。在该系统中由模式匹配模块计算相关度，采取漫游模型来进行后期的持续检索。该系统的最大优点是准确度高，其代价是牺牲了覆盖度，搜索的深度非常有限。

课题研究主要内容、实施方案及创新点：

完成主要内容：

1、设计核心文本舆情爬取算法。

利用该平台采用网络爬虫和模拟访问技术抓取网站上的 Web 内的信息，综合运用语义网、神经网络、模式识别等技术建立适合于进行 Web 舆情爬取的核心算法。利用 Python 实现，该算法可对目标信息进行语意判断的智能化分析，使得爬取系统能够做到智能化检测和分析信息大意，从而提高信息的筛选准确度。

2、数据库设计。

对数据库进行逻辑化管理，并处理网络链接（如网站间的相互链接问题），提高信息分析处理的准确性，并实现信息的完整性分析。对信息分析和判断后，重新呈现页面，作为证据保留在数据库中。

3、报文拦截和跳转提示程序。

对在通信干路上检测用户的报文，使用 Python 中的数据抓包工具对用户的数据包进行监听，如果此数据包中包含色情、暴力、邪教等不良信息或用户要访问的地址为非法网站，则自动丢弃该数据包，阻断用户的访问，或进行网页跳转同时提示用户“网页包含非法内容拒绝访问”，从而从源头上杜绝访问非法网站的目的。

实现方案：

1、选取知乎和百度贴吧为舆情爬取主要网址。

2、用 Python 编程实现对该网站的爬取的分类，录入数据库。

3、提供数据查询和关键字检索的统一管理。

创新点：

1、设计出一个算法能够自动对由网络爬虫抓取到的网页进行分析，从而过滤网页中的不良内容 ，将有用的信息录入数据库。

2、用 Python 编程实现对舆情信息的分类和索引，方便用户查询。

课题进度安排:

本项目拟使用 Python 和 MySQL 数据库来实现系统的设计。以下是各个阶段的任务与计划：

2016 年 2 月 24 日到 2016 年 2 月 28 日完成需求分析阶段

主要对系统的需求和功能进行全面的分析。

2016 年 3 月 1 日到 2016 年 3 月 31 日完成总体设计阶段

主要完成以下任务：概要设计（包括界面）、数据库设计。

2016 年 4 月 1 日到 2016 年 4 月 30 日完成详细设计阶段

主要完成各个功能模块的编码工作。

2016 年 5 月 1 日到 2016 年 5 月 7 日完成系统测试

主要对系统进行单元测试，最终提交设计成果。

2016 年 5 月 8 日到 2016 年 6 月 1 日完成材料整理及撰写报告阶段

主要完成以下任务：整理所有文档材料并归档、撰写毕业设计报告。

主要参考文献:

[1]王艺.《重大突发公共事件的微博舆情监测与引导初探》. 贵州民族学院学报. 2011.

[2]张超.《文本倾向性分析在舆情监控系统中的应用研究》(硕士学位论文). 北京邮电大学. 2008.

[3]莫溢，刘盛华，刘悦，程学旗.《一种相关话题微博信息的筛选规则学习算法》. 中文信息学报. 2012.

[4]陆浩.《网络舆情监测研究与原型实现》. 北京邮电大学. 2009.02

[5]莫溢，刘盛华，刘悦，程学旗.《一种相关话题微博信息的筛选规则学习算法》. 中文信息学报. 2012.

[6]杨涛.《智能信息处理技术在互联网舆情分析中的应用》(硕士学位论文).同济大学.2008.

指导教师意见: 同意开题。 指导教师：董永权
学院意见： 学院（公章）：　　学院领导：