河北工业大学

人工智能与数据科学学院

校内实习报告（ 一 ）

报告题目： 科技信息智能采编系统-调研分析

学　　院： 人工智能与数据科学学院

系（专业）： 软件工程

班　　级： 软件 161 班

学　　号： 164552　164558　164567

学生姓名： 董宇航　石凯峰　梁山

实习单位： 河北工业大学

2019 年　10 月 18

一、 研究背景

无论是科研还是学习人们都需要通过网络去寻找最新的专业信息和新闻动态,但信息的爆炸式增长,也让人们越来越难以从信息海洋中快速获取所需信息。一方面是因为网络信息量与日俱增,且更新速度非常快,需要投入大量的时间进行信息的搜索；另一方面,网络上的信息存在大量重复的现象,且格式非常不规范,更加大了用户搜寻信息的难度。

科技类网站提供了大量的前沿科技信息和实时性的科技政策，对现代科技的发展发挥着重要的作用，也受到越来越多用户的欢迎。但网络信息资源数量庞大，网站缺乏有效管理，获取信息效率低下。用户可以通过搜索引擎检索出大量信息,却不能对信息进行提取、组织和处理,随着信息化的进步,人们对获取信息的要求越来越高,信息搜索也从“通用”进入“个性和智能”。目前市面上已经出现了很多信息采集的工具,这些工具可以在一定程度上满足用户获取信息的需求,但是对信息的处理却不尽人意。

自动采编科技网站上的科技信息是提高信息获取效率和高效利用科技信息的重要手段。设计实现科技信息智能采编系统，利用网络爬虫技术和智能技术对科技信息进行自动采集、分类，有利于提高科研人员信息获取效率。通过对科技信息智能采编系统的设计、实现，为用户提供高效收集、分类、导出科技信息的平台，从而帮助提高科研机构和科研人员获取和利用科技信息的能力。

二、 课题意义

网络爬虫是一种按照一定的规则，自动地抓取万维网信息的程序或者脚本。本课题主要结合物联网产品开发项目实际需求，完成基于 Spring Boot 的科技信息智能采编系统的设计与实现。基于 Spring Boot 的科技信息智能采编系统：通

过 PhantomJs 实现对于动态页面的抓取。使用 JSoup 库和正则表达式，实现网页信息的提取。通过 Kashgari（基于 Keras 的简单而强大的 NLP 框架）和 THUCNews（数据集）训练 CNN 网络，实现对网页文本的分类。结合 python 的 SnowNLP 库，实现对网页文章的摘要获取。结合 python 的 jieba 库，提取网页文章的关键词。通过 Spring Boot 搭建网络平台，整合上述技术，实现该系统。通过本课题的设计与实现，深化了对网络爬虫的认识，提高了综合知识的应用能力、独立思考和动手实践解决工程实际问题的能力。

本课题对提升科研人员或科研机构的科技信息获取和利用效率，使科研人员更加高效开展研究工作具有重要意义。

三、 技术分析

将从系统设计、实现角度，对课题中所需的关键技术进行阐述。

1、PhantomJs

PhantomJS 是一个基于 webkit 的 javascript API。它使用 QtWebKit 作为它核心浏览器的功能，使用 webkit 来编译解释执行 JavaScript 代码。它不仅是个隐形的浏览器，提供了诸如 CSS 选择器、支持 Web 标准、DOM 操作、JSON、html5、Canvas、SVG 等，同时也提供了处理文件 I/O 的操作，从而使你可以向操作系统读写文件等。

2、Spring Boot

SpringBoot 是由 Pivotal 团队在 2013 年开始研发、2014 年 4 月发布第一个版本的全新开源的轻量级框架。它基于 Spring4.0 设计，不仅继承了 Spring 框架原有的优秀特性，而且还通过简化配置来进一步简化了 Spring 应用的整个搭建和开发过程。

3、JSoup

JSoup 是一款 Java 的 HTML 解析器，可直接解析某个 URL 地址、HTML 文本内容。它提供了一套非常省力的 API，可通过 DOM，CSS 以及类似于 jQuery 的操作方法来取出和操作数据。

4、TextRank

TextRank 算法是一种用于文本的基于图的排序算法。其基本思想来源于谷歌的 PageRank 算法，通过把文本分割成若干组成单元(单词、句子)并建立图模型，利用投票机制对文本中的重要成分进行排序，仅利用单篇文档本身的信息即可实现关键词提取、文摘。

5、SnowNLP

SnowNLP 是一个 python 写的类库，可以方便的处理中文文本内容，是受到了 TextBlob 的启发而写的。

6、JPA

JPA 是 Java Persistence API 的简称，中文名 Java 持久层 API，是 JDK 5.0 注解或 XML 描述对象－关系表的映射关系，并将运行期的实体对象持久化到数据库中。

7、TextCNN

将卷积神经网络 CNN 应用到文本分类任务，利用多个不同 size 的 kernel 来提取句子中的关键信息（类似于多窗口大小的 ngram），从而能够更好地捕捉局部相关性。

四、 课题内容

该系统的目标是利用网络爬虫技术和智能技术对科技信息进行自动采集、分类、存储，科技信息包括新闻信息和科技机构网站信息。当用户指定目标网站后，系统开始自动对其进行信息采集、处理、整理、分类最终存入数据库。用户可通过系统页面对已经采集的信息可进行编辑、保存、导出等操作。

所需功能：

1、用户信息管理功能。对用户账号、密码、用户名、头像等基本信息进行维护。用户通过输入邮箱和密码，进行账号注册。用户可以修改密码、找回密码、修改头像、修改背景图等。

2、爬虫策略管理。管理员可以查看、创建、修改、删除爬虫策略。

3、爬虫入口网址管理。管理员可以查看、创建、修改、删除爬虫策略相对应的爬虫入口网址。

4、主题管理。用户可以创建爬虫主题，方便创建爬虫任务。用户可以对创建的主题进行修改、删除等操作。

5、任务管理。用户通过选择系统已定义的爬虫模板，创建爬虫任务。用户可以创建、查看、启动爬虫任务。

6、文本分类功能。系统根据爬取的文章的标题、内容等信息，对文章进行自动分类。

7、关键词提取功能。系统可以对一篇或一组文章进行关键词提取。

8、简报生成功能。系统通过 TextRank 算法，对一篇或一组文章进行摘要的提取，并导出简报。

9、信息管理功能。用户可以对于爬虫任务爬取下来的数据进行查看、修改

等操作。用户可以选择将要导出的数据，并已 Excel 的形式导出下载。用户可以对爬虫获取的信息进行编辑。

10、信息提取功能。用户定义爬虫策略时，可以对网页文章标题、作者、内容、标题等关键信息进行提取，可以对图片等进行处理。

11、信息排重功能。系统需要实现网页抓取排重和发布信息的排重。

12、爬虫爬取功能。系统通过 PhantomJs 实现对动态页面的爬取，以处理网站的反爬虫策略。

五、 系统软硬件环境

1、开发平台：Ubuntu16.04、Windows10

2、开发工具：IntelliJ IDEA、Pycharm、Chrome Dev

3、数据库：MySQL

六、 所需技术

1、使用 TextRank 算法对文章摘要生成、关键词提取。

2、采用 TextCNN 和 THUCNews（数据集），训练文本分类模型，实现网页文章自动分类。

3、采用 MySQL 数据库实现数据存储。

4、采用 JPA 实现数据库高效、快捷的访问和读取。

5、采用 JSoup 和正则表达式对网页信息进行筛选。

6、使用 Java 设计开发 B/S 结构的基于 Spring Boot 的科技信息智能采编平台。

7、使用 MD5 对用户密码等信息进行加密。

参考文献

[1] 胡琼. 农业网站科技信息智能采编系统研究[D]. 华中农业大学, 2010.

[2] 贾自艳. Web 信息智能获取若干关键问题研究[D]. 中国科学院研究生院（计算技术研究所）, 2004.

[3] 邹丽娜. 网络信息采集及智能处理技术研究[D]. 广东工业大学，2012.

[4] 刘飞. 网页信息智能采集与分类的研究与实现[D].河北工业大学，2014.

[5] 苏旋. 分布式网络爬虫技术的研究与实现[D]. 哈尔滨工业大学，2006.

[6] 李舒晨 网络信息采集处理平台的研究[D]. 北京交通大学, 2009.

[7] 彭雅 文本分类算法及其应用研究[D]. 湖南大学, 2004.

[8] 孙茂松，李景阳，郭志芃，赵宇，郑亚斌，司宪策，刘知远. THUCTC：一个高效的中文文本分类工具包. 2016.

[9] 杨君. 基于 Scrapy 技术的数据采集系统的设计与实现[D]. 南京邮电大学,2018.

[10] 宦俊伟. TrenData 数据分析平台数据爬取与处理模块的设计与实现[D]. 南京大学,2014.