**西安邮电大学**

**毕业设计（论文）**

题目：基于NLP的特定新闻博客摘录系统

学院： 计算机学院

专业： 软件工程

班级： 软件1304班

学生姓名： 袁曙光

学号： 04133109

导师姓名： 孟伟君 职称： 讲师

起止时间：2016年12月5日至2017年6月2日

毕业设计（论文）声明书

本人所提交的毕业论文《基于NLP的特定新闻博客摘录系统》是本人在指导教师指导下独立研究、写作的成果，论文中所引用他人的文献、数据、图件、资料均已明确标注；对本文的研究做出重要贡献的个人和集体，均已在文中以明确方式注明并表示感谢。

本人完全理解《西安邮电大学本科毕业设计（论文）管理办法》的各项规定并自愿遵守。

本人深知本声明书的法律责任，违规后果由本人承担。

论文作者签名：

日期： 年 月 日

西安邮电大学本科毕业设计(论文)选题审批表

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 申报人 | 孟伟君 | | 职称 | | 讲师 | | | 学院 | | 计算机学院 | | | |
| 题目名称 | 基于NLP的特定新闻博客摘录系统 | | | | | | | | | | | | |
| 题目来源 | 科研 | √ | | | | | | 教学 | |  | 其它 | |  |
| 题目类型 | 硬件  设计 |  | | 软件  设计 | | √ | | 论文 | |  | 艺术  作品 |  | |
| 题目性质 | 应用研究 | | | √ | | | | 理论研究 | | |  | | |
| 题目  简述 | 现在对于互联网信息爆炸的形势，人们没有充足的的时间阅读和理解所有的信息，但是却不能忽略重要信息，经常要对重要的博客和网络栏目进行持续追踪。 正是基于这一现状，本系统就是要整合提取需要关注的博客或是专栏的文章摘要和关键词，实现信息的初步整合和文章重点提取，防止遗漏一些信息。 本课题实现的功能是实现内容型网页的正文提取，进行中文分词分析，提取关键字，文摘和其他分析功能，对于需要关注的专栏在服务器后台持续追踪，实时更新，并在web端进行展示。 | | | | | | | | | | | | |
| 对学  生知  识与  能力  要求 | 具备机器学习有关知识；理解计算机网络及WEB的原理和技术，掌握数据库、面向对象技术、c/c++开发技术及相关开发工具，熟悉系统建模方法。 具备独立的面向对象的系统分析与设计的能力。 相关开发环境及工具：操作系统不限，建模工具任选一种支持UML2.0标准的即可，推荐Trufun plato或StarUML，编程工具自定，数据库可用mysql 5.6。 | | | | | | | | | | | | |
| 具体  任务  以及  预期  目标 | 选题学生应完成题目简述中功能，预期目标让学生掌握相关技术，学习相关理论知识，具备相关企业级开发能力，以满足企业对人才的需求。 成果形式是软件系统和毕设论文。 | | | | | | | | | | | | |
| 时间  进度 | 2016.12.5-2016.12.17 查阅资料、理解任务、进行调研，完成开题报告 2016.12.18-2016.12.25 系统需求分析  2016.12.26-2017.1.3 系统概要设计  2017.1.4-2017.1.24 系统详细设计  2017.1.25-2017.3.27 系统实现  2017.3.28-2017.4.10 系统测试与完善  2017.4.11-2017.5.22 撰写论文  2017.5.23-2017.6.2 准备答辩 | | | | | | | | | | | | |
| 系（教研室）主任  签字 | 年 月 日 | | | | | | 主管院长  签字 | | 年 月 日 | | | | |

西安邮电大学本科毕业设计（论文）开题报告

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 学生姓名 | 袁曙光 | 学号 | 04133109 | 专业班级 | 软件1304 |
| 指导教师 | 孟伟君 | 题目 | 基于NLP的特定新闻博客摘录系统 | | |
| 选题目的（为什么选该课题）  本课题为应对信息时代信息过载现象，现在人们没有充足的的时间阅读和理解所有的信息，但是又不能忽略重要信息，经常要对重要的博客和网络栏目进行持续追踪，所以本课题的项目任务是对于重点热点信息进行快速归纳，提取主干摘要，然后推送到用户，使用户能够用最短的时间获取对自己有效有用的信息。从而节省时间，提高效率，更好的信息爆炸时代生活学习 。本课题实现的功能是实现内容型网页的正文提取，进行中文分词分析，提取关键字，文摘和其他分析功能，对于需要关注的专栏在服务器后台持续追踪，实时更新，并在web端进行展示。现阶段，新闻聚合系统非常流行，但是主要是根据浏览记录的轨迹进行相关推荐，存在被动性和单一性，对于重点关注对象也无法持续追踪，而且不能用户主动定制，无法及时推送。所以本项目，就是要进行重要信息的自动抓取推送工作，让人们不遗漏任何必要信息，具有很强的优势和便利性。 | | | | | |
| 前期基础（已学课程、掌握的工具，资料积累、软硬件条件等）  1．网页正文进行识别提取，这一个过程有几种办法，可以使用正则表达式提取，可以对HTML文本的DOM结构分析提取正文，可以使用部分新颖方法如正文块文本提取法等，具有成熟方法。 2.文本的分词和提取文摘，提取关键词。这一过程设计NLP的知识，对于中文的分词可以使用第三方词库基础上匹配，对于提取文摘何关键词使用TextPage算法可以较好的实现功能。 3.结果的展示，前端使用html,css,javascript实现网站，php语言操作数据库，在lamp架构下实现。 | | | | | |
| 要研究和解决的问题（做什么）  现在对于互联网信息爆炸的形势，人们没有充足的的时间阅读和理解所有的信息，但是却不能忽略重要信息，经常要对重要的博客和网络栏目进行持续追踪。正是基于这一现状，本系统就是要整合提取需要关注的博客或是专栏的文章摘要和关键词，实现信息的初步整合和文章重点提取，防止遗漏一些信息。 | | | | | |
| 工作思路和方案（怎么做）  本课题的工作方案分成三阶段： 1.学习并完成中文分词，TextPage算法，网页提取算法相关的知识和实现。通过查找资料询问学长老师，完成相关核心算法的实现。预计花费两个月。使用c来完成算法部分的工作。计划时间段：2016.12.5-2017.1.31 2.实现前端展示页面的实现，使项目具有友好的展示界面，便于显示相关的内容。使用html,css,js,php等技术，b/s架构系统，预计花费一个半月。计划时间段：2017.2.1-2017.3.15 3.前后端进行整合，进行系统测试，完成论文撰写。预计花费一个半月。计划时间段：2017.3.15-2017.6.2。 | | | | | |
| 指导教师意见  签字： 年 月 日 | | | | | |

西安邮电大学毕业设计 (论文)成绩评定表

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 学生姓名 | 袁曙光 | 性别 | | 男 | 学号 | 04133109 | 专业  班级 | | 软件工程  1304班 |
| 课题名称 | 基于NLP的特定新闻博客摘录系统 | | | | | | | | |
| 指导  教师  意见 | （从开题论证、论文内容、撰写规范性、学习态度、创新等方面进行考核）  评分（百分制）： 指导教师(签字)：　 年 月 日 | | | | | | | | |
| 评阅  教师  意见 | （从选题、开题论证、论文内容、撰写规范性、创新和预期成果等方面进行考核）  评分（百分制）： 评阅教师(签字)：　 年 月 日 | | | | | | | | |
| 验收  小组  意见 | （从毕业设计质量、准备、操作情况等方面进行考核）  评分（百分制）： 验收教师(签字)：　 年 月 日 | | | | | | | | |
| 答辩  小组  意见 | （从准备、陈述、回答、仪表等方面进行考核）  评分（百分制）： 答辩小组组长(签字)：　 年 月 日 | | | | | | | | |
| 评分比例 | 指导教师评分(30％) 评阅教师评分(20％) 验收小组评分(20％) 答辩小组评分(30％) | | | | | | | | |
| 学生总评  成绩 | 百分制成绩 | |  | | | 等级制成绩 | |  | |
| 答辩委员会意见 | 毕业论文(设计)最终成绩(等级)：  学院答辩委员会主任(签字、学院盖章)：年 月 日 | | | | | | | | |

摘　　要

面对信息时代的信息过载，人们没有充足的时间去阅读和理解所有的信息，但是又不能忽略重要信息，经常要对重要的博客和网络栏目进行持续追踪，所以本课题的项目任务是对于重点热点信息进行快速归纳，提取主干摘要，然后推送给用户，使用户能够用最短的时间获取对自己有效有用的信息。

现阶段，新闻聚合系统非常流行，但是主要是根据浏览记录的轨迹进行相关推荐，存在被动性和单一性，对于重点关注对象也无法持续追踪，而且用户不能主动定制，无法及时推送。所以本课题，就是要进行重要信息的自动抓取推送工作，让人们不遗漏任何必要信息，具有很强的优势和便利性。本项目实现的功能是围绕NLP相关技术实现内容型网页的正文提取，进行中文分词分析，提取关键词，提取文摘和其他分析功能，对于需要关注的专栏在服务器后台持续追踪，实时更新，并在web端进行展示。

论文首先介绍NLP中的中文分词，提取关键词，网页内容提取的国内外研究现状，明确现有技术的情况和局限性，提出了本课题的主要内容和意义。

在提取网页文本方面，使用了基于行块分布函数的算法进行提取，可对正文型网页的内容进行自适应提取，忽略HTML格式和排版情况。在提取关键词和文摘方面，使用TextRank有监督机器学习算法，并辅助BM25算法计算句子相似性，通过投票实现关键词和文摘提取。

最后以实现的角度阐述了基于NLP的特定新闻博客摘录系统的详细设计与实现，如系统流程与实现、数据库表设计、逻辑流程。对于各个子模块，使用逻辑流程图或伪代码算法描述的方法给予说明，然后对系统的各个子模块进行实验，展示其实现效果。

关键词：TextRank；BM25；文摘提取；关键词提取

**ABSTRACT**

In the era of information, people do have limit time to read and understand all the information.but they won’t ignore important information by keeping track of the blog and network program important, so this project is for the key task for the rapid induction hot information and extract the abstract, then the information pushed to the user. The user can get the useful information in the shortest time.

At this stage, news aggregation system is very popular, but is mainly related to recommend the browsing path, there are passive and unitary, the focus cannot keep track of, but not active user customization, unable to timely push. Therefore, the project is to carry out the automatic capture and push of important information, so that people do not miss any necessary information, and has a strong advantage and convenience. The project based nlp to achieve the function of content extraction for Chinese segmentation analysis, keyword extraction and analysis and the need to pay attention to the column in the server keeps track, real-time updates, and display on the web side.

This paper introduces the Chinese word segmentation, extraction of keywords, web content extraction research status and the existing situation and limitations of the technology, and put forward the main content and significance of this topic.

In the extraction of web pages, we use the algorithm based the block distribution function to extract the content of pages, and ignore the HTML format. In the extraction of keywords and abstracts, the TextRank supervised machine learning algorithm is used, and the BM25 algorithm is used to compute sentence similarity.

Finally, from the point of view of implementation, elaborate the detailed design and implementation of system, such as system process and implementation, database table design, logical process. For each sub module, the method is described by logical flow chart or pseudo code algorithm. experiments are carried out on each sub module of the system to demonstrate its effect.

**Key words：**TextRank；BM25；keywords；summary

目　　录

[第一章 引言 1](#_Toc484950154)

[1.1研究背景与意义 1](#_Toc484950155)

[1.2研究历史与现状 1](#_Toc484950156)

[1.3课题主要研究内容 3](#_Toc484950157)

[1.4论文组织结构 3](#_Toc484950158)

[第二章 中文分词方法研究 5](#_Toc484950159)

[2.1中文分词主要算法 5](#_Toc484950160)

[2.2中文分词的难点 6](#_Toc484950161)

[第三章 Web关键词文摘提取关键技术研究 8](#_Toc484950162)

[3.1 TextRank算法提取关键词 8](#_Toc484950163)

[3.2 TextRank算法提取文摘 9](#_Toc484950164)

[3.3 其他方法提取关键词和文摘 10](#_Toc484950165)

[第四章 Web正文提取技术研究 12](#_Toc484950166)

[4.1Web网页特征 12](#_Toc484950167)

[4.2行块函数 12](#_Toc484950168)

[第五章 系统设计与实现 14](#_Toc484950169)

[5.1系统功能 14](#_Toc484950170)

[5.2数据库表结构设计 14](#_Toc484950171)

[5.3功能模块设计与实现 17](#_Toc484950172)

[5.4项目展示 25](#_Toc484950173)

[第六章 总结与展望 28](#_Toc484950174)

[6.1项目工作总结 28](#_Toc484950175)

[6.2研究工作展望 28](#_Toc484950176)

[致谢 29](#_Toc484950177)

[参考文献 30](#_Toc484950178)

第一章　　引言

1.1研究背景与意义

在人类发展的过程中，信息的交流和传递是至关重要的，过去，人们通过阅读报纸来获取有效信息。而对于报纸内容的选择性阅读，是通过标题和摘要，快速获得有效信息和过滤无效信息，筛选出自己愿意阅读的内容。现今，人们更倾向于网络阅读新闻和文章。对于更大数量级的信息，需要正确高效的筛选方法。数量巨大的在线书城和新闻站点占据了人越来越多的时间，内容从军事、技术、笑话、小说到论文，无所不包。所以，有效的信息检索和处理非常要重要，现在对于互联网信息爆炸的形势，人们没有充足的时间阅读和理解所有的信息，但经常要对重要的博客和网络栏目进行持续追踪。正是基于这一现状，本系统就是基于NLP技术，整合提取需要关注的博客或是专栏的文章摘要和关键词，实现信息的初步整合和文章重点提取，防止遗漏一些信息。

当前，人们对于信息的收集主要是靠各大搜索引擎。搜索引擎通过对收录网站的检索，以标题和内容的形式集中返回数十条信息，这种方式对于一般需求的人来说是足够的。但对信息的及时性，侧重性有需要的人来说，一条条去筛选时难以节省时间，无法做到对某些特定网站的关注。如果能够定制化收集特定需求信息，并提供关键词和文摘，就可帮助人们减少在浏览网页正文所花的时间，更快的提取有效信息。

现在人们没有充足的的时间阅读和理解所有的信息，但经常要对重要的博客和网络栏目进行持续追踪，所以本课题的任务是基于NLP相关技术对于重点热点信息进行快速归纳，提取主干摘要，然后推送到用户，使用户能够用最短的时间获取对自己有效有用的信息。从而节省时间，提高效率。现阶段，新闻聚合系统非常流行，但是主要是根据浏览记录的轨迹进行相关推荐，存在被动性和单一性，对于重点关注对象也无法持续追踪，而且不能用户主动定制，无法及时推送。所以本课题，就是要进行重要信息的自动抓取推送工作，让人们不遗漏任何必要信息，具有很强的优势和便利性。

自然语言处理是计算机科学中非常重要的一个领域，它是一个研究人和计算机间通过自然语言进行沟通的学科，包括在文本方面研究检索系统，机器翻译，文本摘要等。我通过研究自然语言处理中的文本摘要，关键词提取，中文分词，正文提取技术，完成本课题。

1.2研究历史与现状

在网页正文提取方面，研究者有许多成熟方法，进行了比较长时间的研究和实验，在这方面，对于英文和中文的提取并没有很多特别的差别。一般来说正文信息提取是基于HTML标准的,通过抓取WEB页面进行解析,形成DOM结构树，通过比对HTML结构或者分析视觉结构来进行分割处理[1]。使用解析DOM结构，来形成正文DOM块和非正文块，聚类并除杂后获取。或者利用一些特殊标签的出现次数来鉴定正文，比如寻找<img>图片标签或者<a>超链接标签。但是这类传统方式提取web正文是有很多的局限。DOM树方法的前提是每个web页面的编写工作是严谨和符合要求的，如果出现少标签现象，就无法解析成功，适配性很差。对于通过视觉信息提取的方法，无法找出完全匹配任何情况的标签出现次数，对于不同网站，需要再次适配。并且在海量web页面数据下，传统方法的耗时是巨大，需要解析HTML页面，难以快速处理。

在提取关键词方面，关键词自动提取是对正文内容进行统计学或者语义学上的分析，在众多候选词组中，选取合适的能够表达文章的词。表达文章主题的关键词是非常重要的，关键词提取技术是信息检索，搜索引擎的的重要工作，并发挥着不可替代的作用。对于提取关键词的研究，国内外都已经做了大量有意义并且实用的工作，Luhn1957年进行了自动引标的工作，到现在已经有50多年的历史。

(1)基于统计的关键词提取方法。Li研究了TFIDF的方法[2]，对词组进行频率统计和在文档库中的出现频率的倒数的计算，并进行相乘运算。EL-Beltgy研究KP-Miner的方法[3]，通过将停用词和标点将正文分割成词序列，运用规则和频率对词进行过滤。一般来说，这类方法注重考虑词频和词出现的位置，重点在高频词和位置比较特殊的词，可能出现与内容无关的高频词，所以会对准确率有影响。

(2)基于机器学习的关键词提取方法。对于PageRank算法改进适应提取关键词的TextRank算法[4]，基于词互相根据距离投票的原理，迭代多次投票找出关键词，是本课题选用的算法。国外对种类方法的研究比较深入，如Turney研发的Genex系统[5]，将关键词提取应用到机器学习方法中，通过遗传和决策树算法解决实际问题，效果明显并且商用。Witten开发KEA软件包[6]，主要使用贝叶斯算法提取关键词，对文本进行预处理和TDIDF特征处理，形成模型，从而获得文本的关键词信息。机器学习方法将关键词提取问题抽象成分类问题，通过训练获取特征，进而通过大量的训练获得预测参数，在真实文本下就可通过模型获得较好的输出，取得关键词。但是也有许多限制，比如样本量不足问题，样本未标注问题，样本标注不全，难以获得完美的预测参数和分类器，机器学习的欠拟合和过拟合问题也需要进行处理，进而影响提取的准确性。

(3)基于语义的关键词提取方法。从人类对语言理解的方向，解析文本获取关键词，也是近年来的研究方向，因为与人类的思维方式相同，会取得意想不到的效果，得到潜在含义。Gonenc提出词汇链的方法[7]，将语义相关的关键词链接起来，分析特征获取关键词与主旨对应。You提出的方法是通过选举核心词扩展成关键词[8]。

在提取文摘方面。自动摘要的发展研究历史比较长，在各个方面都有分厂重要的作用，最早是由IBM的Luhn提出这个想法。他认为文本是由句子组成，文摘就是文本中重要的句子，所以对于文摘的选择，要充分考虑句子中的词语和文本的关系。70年代以前，这方面的研究主要集中在研究统计学的特征，在标题，词，连词介词，位置等等方面找出规律，从而获得文摘。70年代以后，美国GE研究中心的SCSIOR，德国的TOPIC，都是基于知识理解的。97年，微软正式将文摘功能接入office软件，产生了巨大推动效果。

1.3课题主要研究内容

本课题将在自然语言处理（Natural Language Processing,NLP）技术的网页正文提取，关键词提取和文摘提取三个方面进行对比现有技术，并选择相应技术进行实现。

在网页正文提取方面，将对比现有主流中文正文提取方法，比较其特点和适用场景，进行必要的实验，选择基于行块分布函数的正文提取算法进行web页面正文提取。在提取正文关键词方面，对比TFIDF等主流方法，并进行实验，选择TextRank有监督机器学习算法，进行相关理论学习和实验。在提取正文文摘方面，选择TextRank算法并配合BM25句子相似性算法，取得文摘。

1.4论文组织结构

第一章 引言。首先引出本课题研究的意义，通过前期积累和在知网的搜索，了解相关背景和历史过程，针对正文提取，关键词提取，文摘提取这三个方面，介绍其主要研究内容和工作。

第二章 中文分词方法研究。简述现阶段流行有效的中文分词方法，介绍基于词典的分词和基于统计的分词，基于主动学习的分词方法，对进行比较优缺点和应用可能，并介绍现阶段的歧义切分，未登录词收录的难点问题。

第三章 Web关键词文摘提取关键技术研究。详细介绍TextRank这种机器学习有监督方法对于提取关键词和文摘的原理和详细过程，明确其优势和相关的应用。并详细介绍BM25相似性算法在提取文摘中的作用和本身的性质。然后介绍其他提取关键词和文摘的算法。

第四章 Web正文提取技术研究。首先详细介绍行块分布函数的思想和重要定义。

第五章 特定新闻博客摘录系统设计与实现。首先从系统功能，数据库表结构设计，功能模块设计与实现这三方面，从起始到结束流程进行描述，通过图表和截图展示系统的功能。然后对于重要算法进行个别展示，进行网页正文提取实验，中文分词实验，正文关键词提取实验，正文文摘提取实验。最后对项目逻辑流程以图表的形式清晰的展示。

第六章 总结与展望。在最后，对本课题的工作再次回看，发现了许多不足与改进空间，对改进方向进行了探讨，并希望在未来的工作中能够做出完善。

第二章　　中文分词方法研究

不同于英文分词，中文分词需要在无特别分割符如空格的情况下，对句子进行切割，切割出符合人认知的词序列，要按照规范合理进行切分。相对于英文的空格进行自然的切分，中文句中是不存在天然分割，对于词的分割是比较困难的。举个例子。中文“我是一个孩子”和英文“I’m a kid”，英文可以通过空格划分为“im”“a”“kid”，是结果清晰的，而中文的划分是没有确定性，“我”“是”“一个”“孩子”的正确方案是难以简单通过规律得出。

2.1中文分词主要算法

现阶段中文分词主要有三种方法：基于词典，基于统计和基于主动学习的分词方法。

2.1.1基于词典的分词方法

基于词典的方法是比较原始的方法，但是是非常有效的，它又叫做机械分词。即将文本与词库比对，寻找最合适的词序列，最终匹配成功。词典匹配方法有正向，逆向，最长和最短匹配四种方法。在发展过程中，基于这四种基本方法，改进型算法层出不穷。如在2007年，张海营提出的全二分快速自动分词算法这种词典方法[9]，通过为词典构建Hash表和词索引表，运用全二分算法来加速索引过程，提高了其分词的速度。在2012年，曹月雷运用词典与后缀数组相结合的中文分词方法[10]，对于高频和中频词汇，加入后缀数组，加速索引过程，以极高的效率准确提取，最后在词典下完成分词。

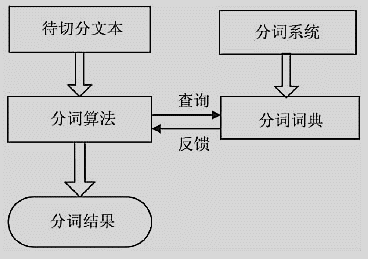


图2.1 词典方法流程图

词典分词的关键在于对词典的合理运用，词典的大小对分词的准确和效率具有决定性的影响,词典的构建结构会影响词的对比和查找。在生产环境中，通过优化词典结构，获得更高效的查询速度，如Hash表，索引表和后缀数组，另外通过高效的查询构建算法。词典分词方法也可以结合其他方法，如结合基于统计的分词方法。

2.1.2基于统计的分词方法

基于统计的分词，基本的分词依据是字之间相似度或者说是紧密度。当待分词句中的字与字之间的连接程度高的时候，超过阈值时，就可被识别为一个词组。最常用的方法是N-gram和最大熵方法。

N-gram认为当前词的出现只和其前一个词出现的概率相关，所有句子出现概率，取决于分词结果概率的乘积。最大熵模型选择在学习概率模型时，所有可能的模型中熵最大的模型是最好的模型；若概率模型需要满足一些约束，则最大熵原理就是在满足已知约束的条件集合中选择熵最大模型。但是会有可能切分出非词汇的组合，如“我的”。该方法的训练集需要标注，不同学科领域的分词，要求的语料库也要适时更换来进行训练。所以，长久以来，研究人员对其方法进行改进，期望获得更好的效率，更好的准确性，在2006年，Wang Xinhao等在Chinese word segmentation with maximum entropy and n.gram language model中提出了把分词问题转化成分类问题[11]，结合N-gram和最大熵方法，进行互补。2010年，Zhang Liyan等[12]，通过实验，构建一个二进制模型，确认在语料库大到一定程度，就不再会影响分词准确性的结论。

基于统计的方法，对于未登录词是友好的，还可以解决新词收录问题，改进词典方法的缺陷。一般来说，统计方法需要使用巨量人工或者机械标注的语料库，需要大量人力和财物消耗。通常来说，将这两种方法结合，可以简化语料库的构建和加快分词的处理过程，减少时间花销，还可以消除歧义，合理切分未登录词。这样的准确率也比两种方法单一使用准确率高。

2.1.3基于主动学习的分词方法

基于主动学习的分词方法是基于机器学习或者是深度学习的方法，是一种新型方法。主动学习方法的一般步骤是，首先选择合适的分类器，输入训练数据，使分类器不断提高其准确率。然后再让实验集输入分类器，获得结果。

常用的机器学习模型主要有感知机模型，马尔科夫模型，条件随机场模型。如隐马尔科夫模型（HMM）是基于时序的有向图概率模型。它包含有三个概率，起始概率，转移概率，发射概率（就是在这个状态获取结果的概率），两个矩阵分别是状态矩阵，观察矩阵。一般输入作为观察序列，输出的分词为状态序列。在大量优质的语料库下，隐马尔科夫模型计算词与词之间转换的概率，和词性产生词的发射概率。

2.2中文分词的难点

2.2.1歧义切分

因为汉语的特点，如结构复杂，使用灵活，词汇多样。在复杂情境下，不要说一个机器难以理解其含义，难以正确分词，就算是熟练运用汉语的人，也会对同一个句子理解产生偏差。所以在分词时，对中文分词产生不同的分词结果，这个问题就是歧义问题。如“这个球的确定制的很好”这个句子就可以切分成“这个/球/的确/定制/的/很好”或者“这个/球的/确定/制/的/很好”两种情况，甚至出现更多种情况。在中文环境中，这种歧义的出现时很难避免的，要通过各种技术和方法解决这个问题。

常见的歧义方式有两种，为交集型和多义组合型。交集型是指在句子中，三个字，前两个和后两个均可组成词语，产生歧义。则中间字为交段，长度为交段长度，交段的数量为链长。如“交并集”这个句子会切分成“交并”，“并集”两个可能的词条，歧义字段长度3，“并”为交段，链长为1。多义组合型是指词拆分单独成词，整体也成词。如“中国人”，可拆分成“中国”“人”或者整体成词“中国人”。

任何高水准的中文分词工具，都需要对歧义现象进行处理，现阶段的解决方法有：基于统计的方法。通过概率计算来确定在上下文中，哪种切分的可能性最大，但是问题在于要考虑数据稀疏问题和实现难度问题。基于规则的方法。语言学家定义出现歧义时的消除歧义依据，通过已定义模型，消除歧义。由于汉语规则的复杂，规则时常出现冲突，难以完全消除歧义，也难以维护。

2.2.2未登录词

词典中未收录的词组，其中有名称，地名，学术名词，微博新出热词，缩略词等等，都称为未登录词。

未登录词的正确分词是中文分词的难点。重要存在以下几个问题：1)中文对于特殊名词没有特殊分类，而英语的特有名词首字母大写。2)新词汇可能本身存在歧义问题，如“袁建国”，可分成“袁”“建国”。3)具有随意性，难以通过已有规则切分。

生产环境中，对于未登录词问题的解决，一般有以下几种方法。第一种，对于可预期的新词进行人工干预，手工或者机器筛选新词加入到词典中，或者机器先行筛选，人工确认，再加入词典。第二种，对于不可预期的未登录词，人工是无法及时添加到词典中，一般就是运用专家归纳的规则收录，或者是通过原料库中统计的情况知识，进行机械识别处理。

一般来说，对于分词准确率来说，未登录词的影响要超过歧义的影响。所以，提高对未登录词的识别处理，对提高分词系统效率，准确度有重要意义。

第三章　　Web关键词文摘提取关键技术研究

TextRank算法是PageRank的变种[13]，是一种基于图模型的无监督机器学习算法，可以用于对文本进行关键词提取或者文摘提取。PageRank是谷歌创始人Larry Page发明，用于对链接进行分析，被用于创立谷歌的搜索引擎算法上，进行网站页面的重要性排名，排序展示搜索结果。PageRank的核心思想是，一个页面被引用的次数越多，引用者权重越高，则这个网页就越重要。

3.1 TextRank算法提取关键词

TextRank是PageRank在文本处理领域的应用算法。它认为将文本分割成单项（词或者是句），每个单项和其它单项相似程度构成图节点与节点之间的权重，进行反复投票，最终收敛，进行排序后输出，形成关键词排名或者文摘排名。

主要过程为：

1)处理文本，形成图模型的节点

2)计算相似度，形成图模型边权重

3)进行迭代投票，获得最终收敛或者到达投票最大次数

4)依据最终得分排序输出结果  
 TextRank的公式为,V代表所有的节点，E代表所有的边。E是V\*V的子集。任意节点的得分计算公式如下：

 (3.1)

WS为权重，初始均为1。d为阻尼系数，用于模拟在网页跳转过程中的情况，在浏览一个网页时，有可能通过页面的超链接跳转，也有可能在url输入框进行跳转，则为非页面超链接跳转，d一般为0.85。In()为所有入链接连接到i节点的集合。Out()为j节点出链接的集合。当误差小到一定程度或者迭代次算到达要求时，停止计算过程。

TextRank算法的计算过程如下：

1)分词，首先利用高效准确的分词库组件，将待提取文本分词。

2)对分词结果进行过滤，删除停用词等无效表达词。

3)创建模型，单词为节点，单词前后相邻的10个左右单词建立边的关系，初始权重为1，形成无向图。

4)由于边的权重为1，更改公式=1，进行迭代计算，直到收敛，误差范围小于规定值，或者迭代次数到达目标要求。

5)根据最终投票得分，提取前N大的关键词进行返回。

这是一般的TextRank计算过程。

3.2 TextRank算法提取文摘

使用TextRank算法提取文摘时，最大差别是图的节点变成了句子，节点之间的权值变成了句子之间的相似性。本课题对于句子之间相似性，使用的是BM25算法，也可以使用词汇重叠方法，字符串核函数，余弦计算，最大共同子序列等方法计算句子之间的相似度，然后代入迭代计算之中的权值W，进行计算，排序降序输出最优结果。其计算过程如下：

1)对文本按照句子进行拆分，但是仍要进行分词，因为在BM25算法计算句子相似度时，词为最小单位。过滤停用词等无效信息。

2)计算句子相似度，BM25算法计算公式如下，它的思想是对查询句进行语素解析，生成语素q。然后，对于每个搜索结果，计算每个语素q与搜索结果的相关性得分，最后，将q相对于结果的相关性得分进行加权求和，从而得到查询句与结果的相关性得分。

 (3.2)

3)创建模型，句子为节点，与所有的句子建立边关系，初始权重为上一步计算的相似性值，形成无向图。

4)进行迭代计算，直到收敛，误差范围小于规定值，或者迭代次数到达目标要求。

5)根据最终投票得分，提取前N大结果的文摘进行返回。

相对于其他方法，TextRank在提取关键词和文摘方面，省去了训练语料库这个步骤，可以适用多领域的文本，无需加入特定规则。缺点是，不考虑局部信息，如词位置等。

3.2.1 BM25算法计算句子相似性

BM25算法的全称是Best Match25算法，是进行文本相似性打分评价的算法，它又名OKapi BM25算法，它来自于OKapi系统用于信息检索而第一次露面。BM25是词袋模型。

BM25公式如下：

 (3.3)

在这个公式中，Q是表示待查询句，q表示待查询句进行分词或者是解析之后的一个单项。d表示所有文档的集合。R为相关性计算，计算单项与文档的相似性。

W权重计算使用IDF算法。公式如下：

 (3.4)

公式中N为文本总数，n为存在单项的文本数量。

通过IDF公式可以看出，当一个单项词在多数文本都存在时，它的权重就相应下降，当时冷僻词时，权重就会上升。

单项与文档集合的相似度计算公式如下：

 (3.5)

 (3.6)

通过更改常数k1，k2和b，可以改变相关性得分的影响因素。通常可以取k1=2，b=0.75。f为单项q在文本中的频率，qf是单项在待查询串的频率。dl是文本d的字数，avgdl文本集合的平均字数。

常数b可以调控文本字数对相关性影响力的能力。当b变大时，文本字数的影响变大，当b变小时，影响转而变小。BM25算法计算相似性，可以根据需求改变常数k1，k2，b的取值，从而改变某些影响因素的影响能力。

3.3 其他方法提取关键词和文摘

除了本课题使用的TextRank算法。还有TF-IDF,朴素贝叶斯，词汇链等方法。

(1)TF-IDF。通过统计词的出现频率，在候选词中提取关键词的算法，优点在于时间开销比较下，在各个领域广泛使用。TF是Term Frequency的缩写，在词频率，IDF是Inverse Document Frequence缩写，代表一个词在全局中的逆权重，出现的越少，则权重越高，可以减少热门词的影响力，增加冷门词的影响力。

1.关键词权重公式：

 (3.7)

TF为词频，IDF为在逆文档频率

2.TF公式，在文档d中关键词w的TF值：

 (3.8)

分子代表关键词w的d文档出现次数。分母代表d中所有词的数量。

3.IDF公式，关键词w的逆文档频率：

 (3.9)

文档共有N个，出现w关键词的文档数量为+1，加一是因为防止出现分母为零的情况。

TF-IDF公式的含义就是在所有文档中，这个词越冷门，则赋予的权重越大，和别的词相比具有区分性，一个词出现次数多，则权重大。

(2)朴素贝叶斯。

朴素贝叶斯是比较基本的有监督类型的机器学习方法。可以将中文关键词提取问题转变为分类问题。所有的候选词在分类器作用下，用是否是关键词这个标准来评判。它的基本原理是贝叶斯概率，它的处理需要两个过程，首先是输入标注后的训练集进行训练，获取参数信息。第二步是输入实验集，完成相关分词任务。相比于TF-IDF方法，朴素贝叶斯的精度有很大提高。但缺陷在于训练集的获取耗时巨大，需要想办法标注出训练集中的关键词或者非关键词。

(3)词汇链。词汇链模型是对文本中的词汇衔接（lexical cohesion）关系进行建模的文本表示模型。目的是表达出文本中复杂的语义信息。词汇连接关系最早是英语语言学家M.A.K.Halliday和R.Hasan提出[14]，明确指出文章中的词的出现并不是随机的，而是与主题和事件有关的。通过构建词汇链，可以确定歧义词的正确切分，对于理解文本有很大的帮助。词汇链是一个文本主题的词链，共同构成文本表达的主旨。

第四章　　Web正文提取技术研究

在基于NLP的特定新闻博客摘录系统中，为了提取关键词和文摘，需要提前将网页正文提取出来，用于进行预处理，如分词，计算句子相似性等。所以在这个课题中，正文提取是至关重要的，提取的纯度进而影响提取关键词和文摘的情况。由于web文本不同于纯文字信息，存在较多的标签结构，较准确的识别正文是比较困难的，在本课题中选用了基于行块分布函数的通用网页正文抽取方法。

## 4.1Web网页特征

现在HTML已经更新至HTML第五代标准，是网页编写领域的霸主语言，它是一种标记性语言，它是通过用户端的浏览器解析展示给用户，所以服务端只需传递信息，减轻了服务端的压力。通过尖括号来描述内容的特性或者类型，尖括号一般成对出现，将内容放在这对尖括号中间，一般后边的尖括号都会有斜杠用于标记结束描述，如“<h1>内容</ h1>”。所有的网页正文均被一对尖括号包围，所以对于其提取产生了难度。

## 4.2行块函数

本课题选用基于行块分布函数的算法，为哈尔滨工业大学陈鑫提出，该算法开源并发布于google code。

如果单纯考虑提取某一个网站的正文，是完全可以以特定规则的正则表达式引擎或者是通过解析DOM来取得正文。但是当更换另一个站点时，这些方法的规则也要随之改变，没有任何通用性。而基于行块分布函数的正文提取算法，是可以忽略Web页面的规则，通过统计学特征，完成提取工作。

它可以在时间复杂度O(n)内抽出正文。它是根据文本密度和起始结束行块的字数来确定正文区。这个算法认为内容型网页字符密度最大的区域肯定是正文区，正文区行块长度不会太短。除杂后，将正文做行块处理。在图上表示为横坐标为行，纵坐标为字数，问题就转化为求字数的骤升骤降点，正文就分布在两点之间的区域内。正文区域需要4个条件：

1)起始点字数需要超过一个阈值。

2)在正文区域每一行块不能出现连续3行的空行。

3)结束点以及其后紧密连续点字数为零。

4)正文区域存在最多行块字数的行。

它的工作过程如下：

1)根据url提取全部HTML信息。

2)进行预处理，去除js脚本文件，css文件和特殊字符。

3)去除html标签，形成纯文本。

4)求出行块分布函数，找出骤升骤降点，提取范围内文字。

它的优点是无需考虑HTML结构，就算结构不全也不影响，因为会当作杂质去除，无需建立dom树模型，而且只需要扫描一次，而正则表达式引擎需要多次扫描。

第五章　　系统设计与实现

本课题是基于NLP的特定新闻博客摘录系统，可以根据用户需求，定制追踪博客或者是新闻板块内容，自动推送关键词和文摘供用户筛选。项目可分解成算法模块和前端展示模块两大类。算法模块主要由正文提取模块、TextRank提取关键词模块、TextRank文摘提取模块、中文分词模块、BM25相似性计算模块、服务器socket通信模块组成，使用c语言编写。前端展示模块使用php，html，css，javascript语言完成，数据库使用mysql，操作系统为ubuntu16.10，内存大小8G，硬盘大小500G，开发工具使用VIM等。

5.1系统功能

本课题主要实现web文本正文提取并进行关键词推荐和文摘提取功能，系统流程图如图5.1所示，共有以下主要功能。

1)正文提取，对于内容型网页提取正文。

2)专栏提取，对于追踪的专栏信息进行提取。包括专栏作者，专栏文章数量等信息。

3)中文分词，用于关键词推荐和文摘提取中的相似性计算。

4)通过TextRank算法获得前N个关键词推荐。

5)通过TextRank算法获得前N个文摘推荐。

6)通过socket服务器编程，将计算信息传输到前端，保存到数据库并展示。

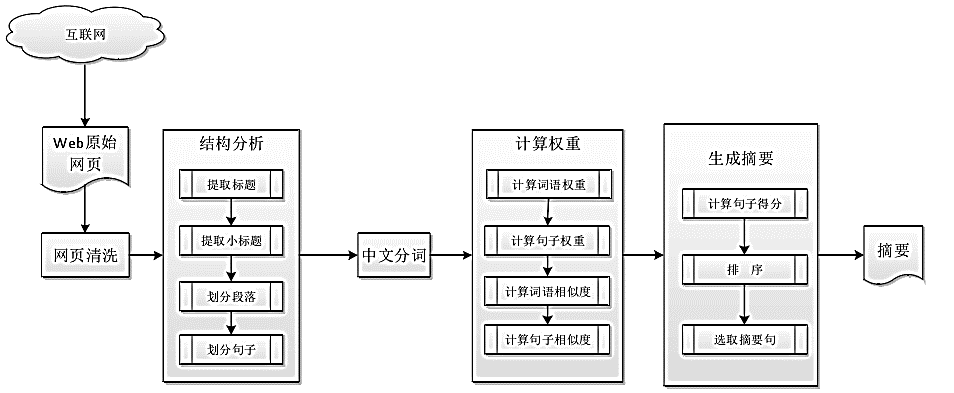


图5.1 系统流程图

5.2数据库表结构设计

此节展示数据库表的结构设计，共需要六个表，分别为用户信息表，专栏信息表，文章信息表，评论信息表，用户专栏连接信息表，用户文章连接信息表。

1)用户信息表结构

用户信息表存放注册用户的用户名，密码，邮箱等必要信息。结构如表5.1所示。

表5.1 用户信息表结构

|  |  |  |  |
| --- | --- | --- | --- |
| 字段名 | 类型 | 主键 | 备注 |
| ID | int | PK | 用户编号 |
| Name | varchar | NO | 用户名 |
| Password | varchar | NO | 用户密码 |
| Email | varchar | NO | 用户注册邮箱 |

2)专栏信息表结构

专栏信息表用于存放追踪的专栏的信息，如专栏名称，作者，更新时间，描述等等。结构如表5.2所示。

表5.2 专栏信息表结构

|  |  |  |  |
| --- | --- | --- | --- |
| 字段名 | 类型 | 主键 | 备注 |
| ID | int | PK | 专栏编号 |
| Title | varchar | NO | 专栏名称 |
| Url | varchar | NO | 地址 |
| Author | varchar | NO | 作者 |
| Others | varchar | NO | 其他信息 |
| Desc | varchar | NO | 描述 |
| Ctime | datetime | NO | 创建时间 |
| Stime | datetime | NO | 最近更新时间 |
| Comment | varchar | NO | 备注 |

3)文章信息表结构

存放文章题目，内容，所属专栏，url地址，关键词，文摘等内容。结构如表5.3所示。

4)评论信息表结构

评价信息表用于记录用户对于某一个文章的笔记随想，包括标题，内容，时间，所属用户编号，所属文章编号。结构如表5.4所示。

5)用户专栏连接信息表结构

用于建立用户和专栏的连接关系。结构如表5.5所示。

表5.3 文章信息表结构

|  |  |  |  |
| --- | --- | --- | --- |
| 字段名 | 类型 | 主键 | 备注 |
| ID | int | PK | 文章编号 |
| Gid | int | NO | 所属专栏编号 |
| Title | varchar | NO | 题目 |
| Ocontent | varchar | NO | 原文 |
| Ncontent | varchar | NO | 正文 |
| Url | varchar | NO | 地址 |
| Others | varchar | NO | 其他 |
| Otime | datetime | NO | 文章发表时间 |
| Ntime | datetime | NO | 文章抓取时间 |
| Comments | varchar | NO | 评论 |
| Keywords | varchar | NO | 关键词 |
| Abstract | text | NO | 文摘 |

表5.4 评论信息表结构

|  |  |  |  |
| --- | --- | --- | --- |
| 字段名 | 类型 | 主键 | 备注 |
| ID | int | PK | 评论编号 |
| Aid | int | NO | 文章编号 |
| Uid | int | NO | 用户编号 |
| Title | varchar | NO | 评论名称 |
| Time | datetime | NO | 时间 |
| content | varchar | NO | 内容 |

表5.5 用户专栏连接信息表结构

|  |  |  |  |
| --- | --- | --- | --- |
| 字段名 | 类型 | 主键 | 备注 |
| ID | int | PK | 编号 |
| Uid | int | NO | 用户表编号 |
| Gid | int | NO | 专栏表编号 |

6)用户文章连接信息表结构

用于建立用户和文章连接关系。结构如表5.6所示。

表5.6 用户文章连接信息表结构

|  |  |  |  |
| --- | --- | --- | --- |
| 字段名 | 类型 | 主键 | 备注 |
| ID | int | PK | 编号 |
| Uid | int | NO | 用户表编号 |
| Gid | int | NO | 专栏表编号 |
| Aid | int | NO | 文章表编号 |

5.3功能模块设计与实现

5.3.1网页正文提取

当系统对某专栏信息进行了提取，并保存数据库后，下一步就应该运用基于行块函数的算法进行正文提取。具体步骤如下：

1)使用开源库提取专栏必要信息。

2)将专栏文章url地址提取，如存在分页则递归完成任务。

3)前端将提取的url传递到后台，使用基于行块函数的算法进行正文提取工作。

4)将正文传递给关键词提取模块和文摘提取模块。

算法实现代码描述如下：

算法名称：基于行块函数的算法

输入：去除杂质后的web页面内容

输出：正文

Void split(string preString) {

int pos=0;

int cur=0;

Pcre p("\\s+");

while((cur=preString.find('\n',pos))!=string::npos){

if (cur-pos-1>1) {

vString.push\_back(p.replace((preString.substr(pos,cur-pos)),""));

}

pos=cur+1;

}

maxline=vString.size();

for(int i=0;i<maxline-blankSize;i++){

int wordsNum=0;

for(int j=i;j<i+blankSize;j++){

wordsNum+=vString[j].length();

}

sizeLines[i]=wordsNum;

}

}

string TextExtract::parse(){

string preString=this->preProcess();

this->split(preString);

int starts = -1;

int ends= -1;

int flagLine[1000];

int maxFlagLine=0;

int curFlagLine=0;

bool boolstart = false, boolend = false;

string res;

for (int i = 0; i < maxline-blankSize; i++) {

if (sizeLines[i] > threshold && ! boolstart) {

if (sizeLines[i+1] != 0

|| sizeLines[i+2] != 0

|| sizeLines[i+3] != 0) {

boolstart = true;

starts = i;

continue;

}

}

if (boolstart) {

if (sizeLines[i] == 0

|| sizeLines[i+1] == 0) {

ends = i;

boolend = true;

}

}

if (boolend) {

for (int ii = starts; ii <= ends; ii++) {

if (sizeLines[ii] < 25){ continue;}

flagLine[maxFlagLine++]=ii;

}

for (; curFlagLine < maxFlagLine; curFlagLine++) {

res.append(vString.at(flagLine[curFlagLine]));

res.push\_back('\n');

}

boolstart = boolend = false;

}

}

return res;

}

正文提取实验结果如图5.2所示。

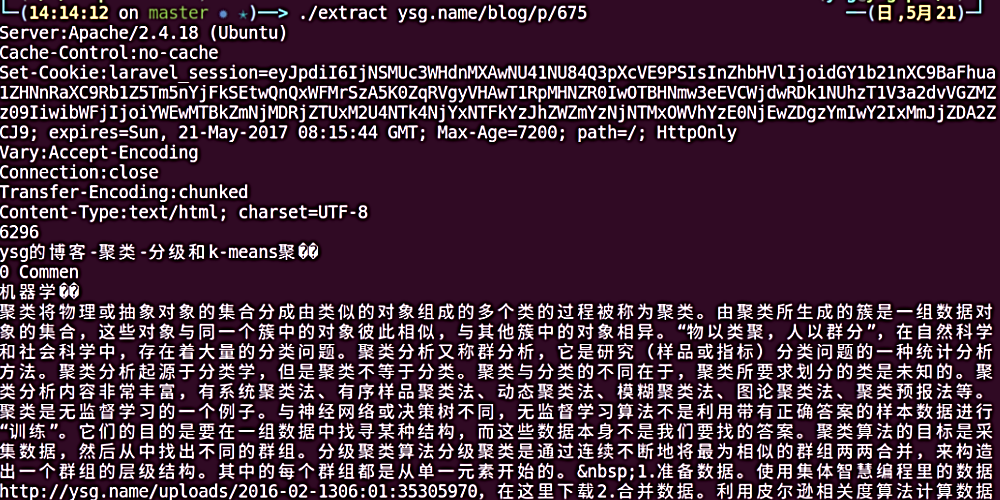


图5.2 正文提取结果展示图

5.3.2中文分词

中文分词在这个项目中有着极大的作用，是关键词提取和文摘提取的前置工程，使用scws开源中文文词库。为了防止开源库词库被多次调用加载占用内存，所以使用单例模式来防止被无意重复加载。

算法实现代码描述如下：

算法名称：单例模式加载scws分词库

输入：空

输出：空

static Participle\* GetInstance(){

selfLocker.lock(); //加锁

if (instance\_==NULL) {

instance\_=new Participle(); //单例模式返回实例

}

selfLocker.unlock(); //解锁

return instance\_;

}

测试文本为：新华社北京4月5日电外交部发言人华春莹5日宣布：应俄罗斯、爱沙尼亚、斯洛文尼亚、阿尔巴尼亚、哈萨克斯坦五国政府邀请，国务院副总理张高丽将于4月11日至19日赴俄罗斯举行中俄投资合作委员会第四次会议和中俄能源合作委员会双方主席会晤，访问爱沙尼亚、斯洛文尼亚、阿尔巴尼亚，赴哈萨克斯坦举行中哈合作委员会第八次会议。

分词实验测试结果如图5.3所示。

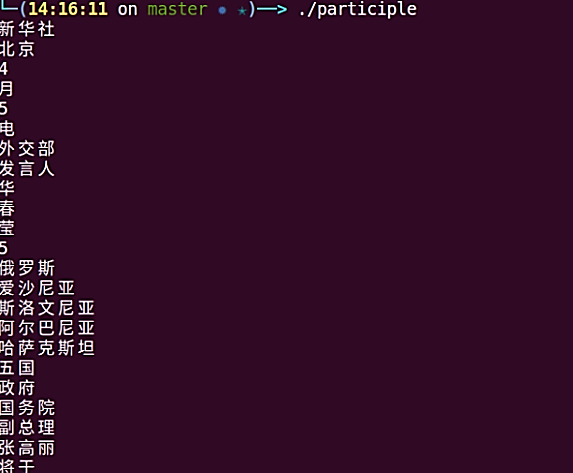


图5.3 分词结果展示图

5.3.3正文关键词提取

正文关键词算法使用TextRank算法，在这个项目中设置200次迭代计算，当处在大批量文本同时需要提取关键词时需要花费一些时间，需要需要分词后的结果作为输入。

算法实现代码描述如下：

算法名称：正文关键词提取算法TextRank

输入：正文提取模块提取结果

输出：排序后关键词结果

vector<pair<string,double>> process(){

int n=text.size();

int cnt=0;

for (int i = 0; i < n; i++) {

set<string> tmpSet;

string curString=text[i];

if (m.find(curString)!=m.end()) {

tmpSet=m[curString];

}

for (int j = 1; j < 5; j++) {

if (i-j>=0) {

tmpSet.insert(text[i-j]);

}

if (i+j<=n-1) {

tmpSet.insert(text[i+j]);

}

}

m[curString]=tmpSet;

}

map<string,double> score;

for (auto it=m.begin();it!=m.end();it++) {

score[it->first]=1;

}

for (size\_t i = 0; i < max\_iter; i++) {

map<string,double> tmpScore;

double maxdiff=0;

for (auto it:m) {

tmpScore[it.first]=1-d;

set<string> tmpSet=it.second;

for(auto itother:tmpSet){

int ss=m[itother].size();

if (it.first==itother||ss==0) {

continue;

}

tmpScore[it.first]=tmpScore[it.first]+d/ss\*score[itother];

} maxdiff=max(maxdiff,abs(tmpScore[it.first]-score[it.first]));

}

score=tmpScore;

if (maxdiff<=diff) {

break;

}

}

vector<pair<string,double>> vecs;//结果

for (auto it = score.begin();it !=score.end();it ++) {

vecs.push\_back(make\_pair(it->first,it->second));

}

sort(vecs.begin(),vecs.end(),CmpByValue());

return vecs;

}

测试文本为：

北京时间11月26日，ESPN记者DaveMcMenamin报道，多处消息源称，客场99-103不敌猛龙之后，骑士召开了一场只有球员参加的内部会议。勒布朗-詹姆斯和詹姆斯-琼斯提到了球队本赛季不稳定的表现。今天输给猛龙是骑士本赛季第4场败仗，而且4场失利均是在客场，这让球队的话语权领袖们质疑球队的韧性。第三节比赛还剩3分46秒至第四节比赛还剩2分01秒期间，骑士被对手打了一波16-31的比分，被对手完成逆转。更糟糕的是骑士的防守，他们第四节送给对手3次加罚机会（下半场6次），赔上犯规都防不住对方进球。如果要犯规，我们需要让它们奏效，”骑士内线特里斯坦-汤普森说，我不是说要伤害任何人，但是你不能给人送加罚。这是不能接受的。”

正文关键词提取实验测试结果如图5.4所示。

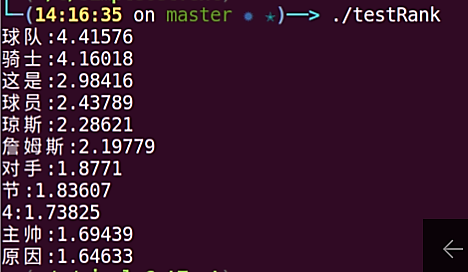


图5.4 正文关键词提取展示图

5.3.4正文文摘提取

正文文摘提取算法使用TextRank算法，并使用BM25算法计算句子之间的相似度。在这个项目中设置150次迭代计算，当处在大批量文本同时需要提取文摘时需要花费一些时间。

算法实现代码描述如下：

算法名称：正文文摘提取算法TextRank

输入：正文提取模块提取结果

输出：排序后文摘结果

vector<pair<string,double>> TextSummary::getTopSentence()

{

int cnt = 0;

for (;cnt<D;cnt++)

{

weight\_sum.push\_back(sumall(weight[cnt]) - weight[cnt][cnt]); // 减掉自己，自己跟自己肯定最相似

vertex.push\_back(1.0);

}

for (int \_ = 0; \_ < max\_iter; ++\_)

{

vector<double> m;

double max\_diff = 0;

for (int i = 0; i < D; ++i)

{

m.push\_back(1 - d);

for (int j = 0; j < D; ++j)

{

if (j == i || weight\_sum[j] == 0) continue;

m[i] += (d \* weight[j][i] / weight\_sum[j] \* vertex[j]);

}

double diff = abs(m[i] - vertex[i]);

if (diff > max\_diff)

{

max\_diff = diff;

}

}

vertex = m;

if (max\_diff <= min\_diff) break;

}

for (int i=0;i<vertex.size();i++) {

vecs.push\_back(make\_pair(sentences[i],vertex[i]));

}

sort(vecs.begin(),vecs.end(),CmpByValue());

for (auto it=vecs.begin();it!=vecs.end();it++) {

cout<<it->first<<":"<<it->second<<endl;

}

return vecs;

}

double TextSummary::sumall(vector<double> arr)

{

double total = 0;

for (double v : arr)

{

total += v;

}

return total;

}

部分测试文本为：北京时间11月26日，ESPN记者DaveMcMenamin报道，多处消息源称，客场99-103不敌猛龙之后，骑士召开了一场只有球员参加的内部会议。勒布朗-詹姆斯和詹姆斯-琼斯提到了球队本赛季不稳定的表现。今天输给猛龙是骑士本赛季第4场败仗，而且4场失利均是在客场，这让球队的话语权领袖们质疑球队的韧性。第三节比赛还剩3分46秒至第四节比赛还剩2分01秒期间，骑士被对手打了一波16-31的比分，被对手完成逆转。更糟糕的是骑士的防守，他们第四节送给对手3次加罚机会（下半场6次），赔上犯规都防不住对方进球。如果要犯规，我们需要让它们奏效。”

正文文摘提取实验测试结果如图5.5所示。

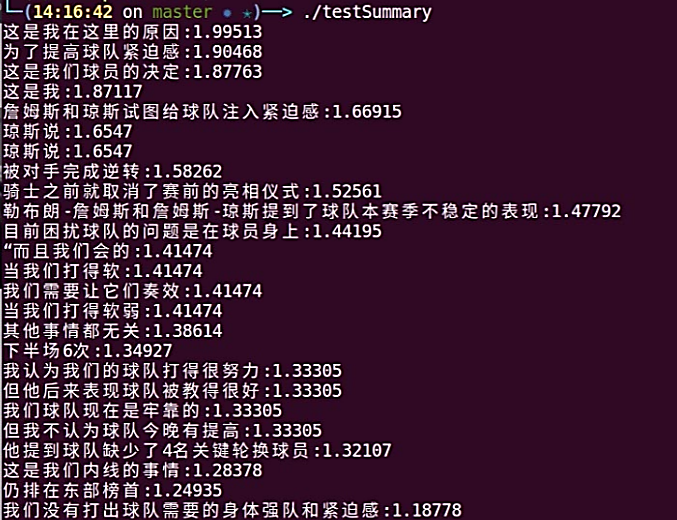


图5.5 正文文摘提取展示图

5.4项目展示

项目登陆页面，为项目入口，提供注册用户进入系统的功能。项目登陆页面如图5.6所示。

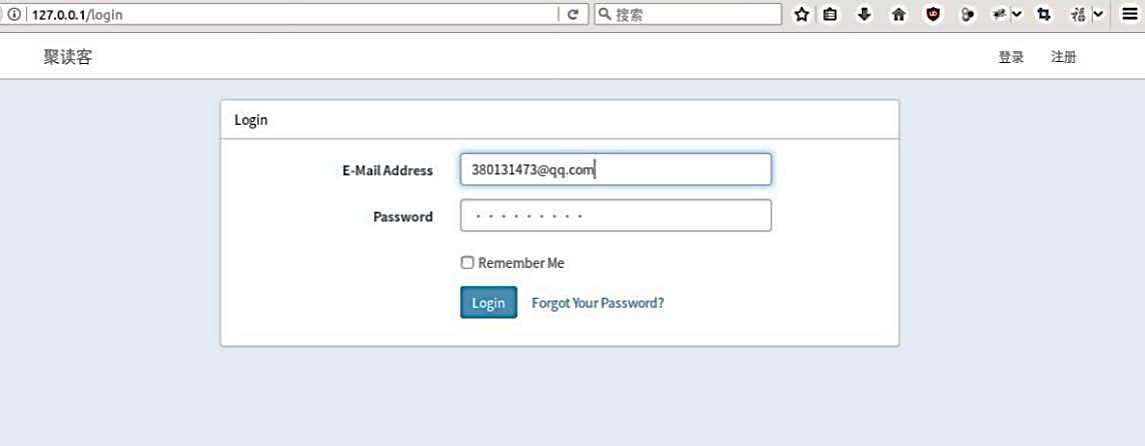


图5.6 登陆页面展示图

用户注册页面，提供新用户注册的功能，注册后返回主页面。用户注册展示图如图5.7所示。

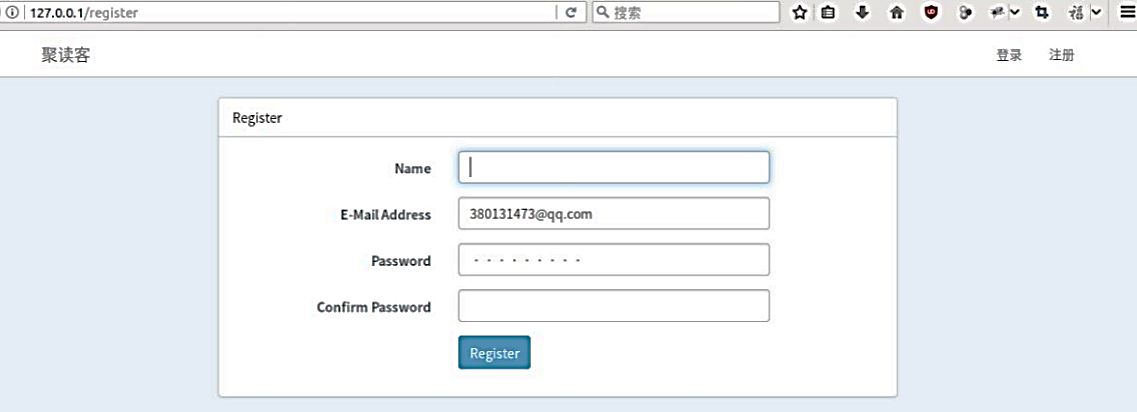


图5.7 用户注册展示图

主页面。提供添加专栏功能，删除专栏功能，更新专栏文章功能。在检索方面提供按关键词，文摘两种索引功能。页面展示文章关键词和文摘。主页展示图如图5.8所示

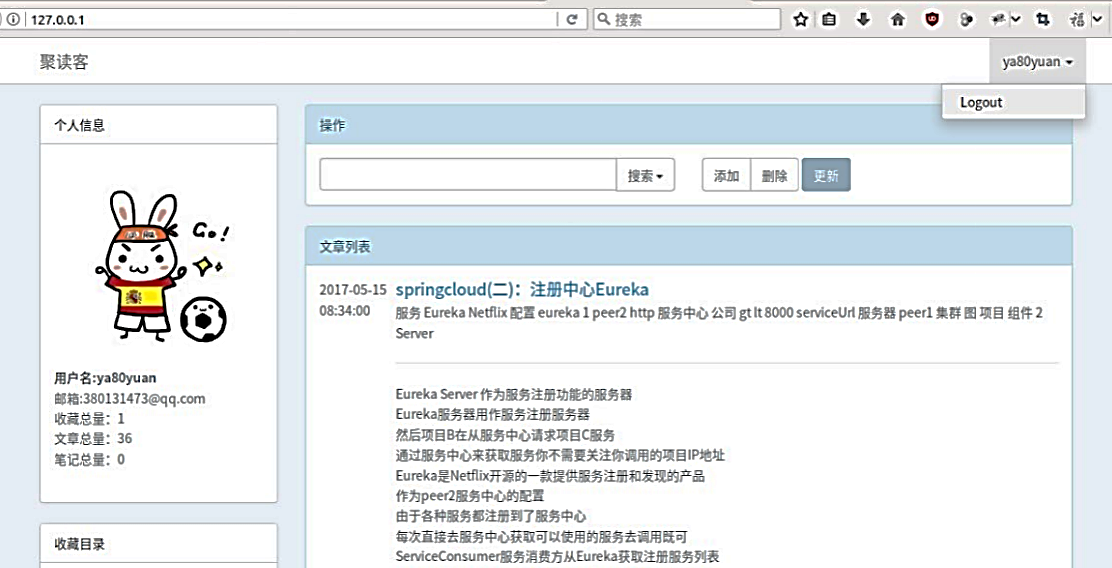


图5.8 主页面展示图

文章内容页。提供原文展示功能，当用户因为提供的关键词或文摘而对文章感兴趣时，点击可查看详情。文章内容页展示图如图5.9所示。



图5.9 文章内容页展示图

第六章　　总结与展望

6.1项目工作总结

本文介绍研究背景，相关技术和实现过程，本课题实现了基于NLP的特定新闻博客摘录系统。详细阐述了了摘录系统对于网页正文提取，中文分词，关键词提取和文摘摘取的过程，各个模块紧密协同完成了总体任务。课题的主要工作分成以下几个板块：

1)中文分词方法研究。简述现阶段流行有效的中文分词方法，介绍基于词典的分词和基于统计的分词，基于主动学习的分词方法，对进行比较优缺点和应用可能，并介绍现阶段的歧义切分，未登录词收录的难点问题。由于分词的巨大工作量，本课题没有从零开始做分词系统，而是使用了开源分词库。

2)详细介绍TextRank这种机器学习有监督方法对于提取关键词和文摘的原理和详细过程，明确其优势和相关的应用，并最终进行了实现。并详细介绍BM25相似性算法在提取文摘中的作用和本身的性质。

3)Web正文提取技术研究。首先详细介绍行块分布函数的思想和重要定义，并对比其他方法，最终在课题中选用此算法，并用c语言进行了实现。

6.2研究工作展望

本课题基于NLP的特定新闻博客摘录系统完成了正文提取模块，分词模块，关键词提取模块，文摘提取模块，需要每个模块紧密协同工作，在编写程序的过程中，我也发现了使用算法和结构的问题，希望在以后的工作学习中，有所突破。问题如下。

1)使用基于行块函数的正文提取算法，对于内容型网页比较好，但是由于其提取是基于正文的分布，但是如果正文字数较少的话，难以正确提取。而且对于图片等非字符信息非常不友好，一般在第一步除杂当中就被去除。而且本来具有的格式被消除，只剩下正文，缺乏应有的人机设计界面。所以改进可以考虑对于某些特定结构，如图片和表格等提前提取，减少有效信息的流失。

2)在分词中，由于代码编写能力不足，使用的开源库完成工作，以后将亲自编写相关程序代码，彻底理解学会中分分词。将会使用隐马尔科夫模型完成这项工作。

3)在使用TextRank算法完成提取关键词和文摘的过程中，当提取文本字数过小时，提取效果较差，因为TextRank算法只需要自身词句切分投票，当文本较小时不能体现关键词代表性。所以在以后工作中，当文本较小时，可以引入TF-IDF方法，以适应各种规模的文本情况。

致　　谢

在这里，首先要感谢我的导师孟伟君老师，与孟老师初识于大一时去北京参加蓝桥杯的火车上，当时老师放弃了卧铺，跟我们一众同学坐硬座。老师时时刻刻关心同学，在学习上为我们指明方向，在教学中严谨扎实。在我因故退出比赛时，并没有苛责反而是关心我，让我能全心投入学习之中，在整个大学期间，孟老师对我的影响非常深厚，学习到了很多宝贵精神和知识。在考研期间，多次关心我的情况，最终我也如愿考入了中国科学院信息工程研究所，这也是老师的栽培的成果，我非常感谢孟老师四年的帮助！

感谢西安邮电大学计算机学院的各位老师。黄茹老师，王小银老师，张德慧老师都给予了我大量的帮助，在我的本科学习中，感受到了老师们的敬业精神和关爱同学的责任心。

感谢大学四年给我留下美好珍贵回忆的同学，感谢西安邮电大学给我珍贵的学习进步的机会，这四年是我思想成熟，习惯养成的时候，也是我真正发展自我的时光，我会永远记得这个地方。

最后，我想感谢我的父母，生我养我，并无时无刻不在关注我的成长，在我成长的道路上为我披荆斩棘，提供养分，为我考虑所有，以后成才定当报答！

参考文献

[1] 刘秉权,王喻红,葛冬梅,李佳.基于结构树解析的网页正文抽取方法[A].龙江省计算机学会2007年学术交流年会论文集[C].2010:14-22.

[2] Sinoara RA,Sundermann CV,Marcacini RM,etal.Named entities as privileged information for hierarchical text clustering. International Database Engineering.&.Applications Symposium.2014:27-33+34.

[3] Samhaa R.El-Beltagy,Ahmed Rafea.KP-Miner:A keyphrase extraction system forEnglish and Arabic documents[J].Information Systems.2010:53-62.

[4] Mihalcea R,Tarau P.TextRank:Bringing Order into Texts.Proceedings of Empirical Methods in Natural Language Processing.2014:39-48.

[5] LI Xing-hua,WU Xin-dong,HU Xue-gang,et al.Keyword extraction based on lexical chains and word co-occurrence for Chinese news Web pages.Proceedings of the 2008IEEE International Conference on Data Mining Workshops.2010:191-200.

[6] Witten I H,Paynter GW,Frank E,et al. KEA:Practical automatic keyphrase extraction.Proceedings of the 4th ACM Conference on Digital Libraries.2012:25-31.

[7] Gonenc Ercan,Ilyas Cicekli.Using lexical chains for keyword extraction[J].Information Processing and Management.2012(6):48-61.

[8] Wei You,Dominique Fontaine,Jean-Paul Barthes.Automatic Keyphrase Extraction with a Refined Candidate Set. Proceedings of the 2009 IEEE WIC ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology,Workshops.2010:83-95.

[9] 火善栋.用哈希算法和二分法实现中文词典的快速查找.计算机时代.2014(09):32-48.

[10]曹月雷,纪文彦,贾斌.词典与后缀数组相结合的中文分词方法[J].硅谷.2012(21):65-78.

[11]Ma,G,Li,X,Rayner,K.Word segmentation of overlapping ambiguous sstrings during Chinese reading. Journal of Experimental Psychology:Humans Perception and Performance.2014:45-76.

[12]Chen Hong.Research on Chinese segmentation algorithm based on Hadoop cloud platform[A].Advances in Computer Science Research[C].2015:23-54.

[13]纪文倩,李舟军,巢文涵,陈小明.一种基于LexRank算法的改进的自动文摘系统[J].计算机科学.2010(05):65-83.

[14]纪文倩,李舟军,巢文涵,陈小明.基于网页正文逻辑段落和长句提取的网页去重算法[J].图书情报研究.2012(02):44-62.