电 子 科 技 大 学

UNIVERSITY OF ELECTRONIC SCIENCE AND TECHNOLOGY OF CHINA

专业学位硕士学位论文

**MASTER THESIS FOR PROFESSIONAL DEGREE**



论文题目 **智能Web新闻文本采集方法研究**

专业学位类别 **学术硕士**

学 号 **201321010806**

作 者 姓 名 **陈西安**

指 导 教 师 **胡光岷 教授**

分类号 密级

UDC注1

学 位 论 文

**智能Web新闻文本采集方法研究**

（题名和副题名）

**陈西安**

（作者姓名）

指导教师 **胡光岷 教 授**

**电子科技大学 成 都**

（姓名、职称、单位名称）

申请学位级别 **硕士**  专业学位类别 **学术硕士**

工程领域名称 **通信与信息系统**

提交论文日期 论文答辩日期

学位授予单位和日期 **电子科技大学 年 月 日**

答辩委员会主席

评阅人

|  |
| --- |
| **LARGE SCALE POINT CLOUD SURFACE RECONSTRUCTION RESEARH AND IMPLEMENTATION FACING THE COMPLEX CONSTRAINTS** |

A Master Thesis Submitted to

University of Electronic Science and Technology of China

Major:  **Electronic and Communication Engineering**

Author: **Yang** **Shuo**

Advisor: **Lu Cai**

School: **School of Communication & Information**

**Engineering**

独创性声明

本人声明所呈交的学位论文是本人在导师指导下进行的研究工作及取得的研究成果。据我所知，除了文中特别加以标注和致谢的地方外，论文中不包含其他人已经发表或撰写过的研究成果，也不包含为获得电子科技大学或其它教育机构的学位或证书而使用过的材料。与我一同工作的同志对本研究所做的任何贡献均已在论文中作了明确的说明并表示谢意。

作者签名： 日期： 年 月 日

论文使用授权

本学位论文作者完全了解电子科技大学有关保留、使用学位论文的规定，有权保留并向国家有关部门或机构送交论文的复印件和磁盘，允许论文被查阅和借阅。本人授权电子科技大学可以将学位论文的全部或部分内容编入有关数据库进行检索，可以采用影印、缩印或扫描等复制手段保存、汇编学位论文。

（保密的学位论文在解密后应遵守此规定）

作者签名： 导师签名：

日期： 年 月 日

摘 要

随着网络的飞速发展,互联网成为大量信息的载体,如何有效地采集并利用这些信息成为一个巨大的挑战，通用搜索引擎己经不能满足人们对个性化信息检索服务日益增长的需要。近年来,各类大大小小的搜索引擎应运而生,以提供分类更细致精确、数据更深入、更新更及时的因特网搜索服务，在搜索引擎的背后是无时无刻不在运行着的网络蜘蛛爬虫系统。本文将对基于聚焦网络爬虫的智能Web新闻采集方法进行研究，智能新闻采集主要面临的问题在于：1、蜘蛛爬虫策；2、非新闻网页的识别；3、网页正文提取，这三个问题将会直接影响新闻采集的优劣。

本文着重围绕新闻采集系统中非新闻网页识别算法和新闻网页文本的智能提取方面展开，本文具体工作如下：

（1）在新闻采集系统中，采用聚焦爬虫策略，其语义分析过程中不简单依赖于关键字，而是通过对文本的各关键词的权重分析，通过计算文本相似性进而确定爬虫方向。

（2）提出基于文本标签特征挖掘的网页正文提取方法。该方法的中心思想基于以下特点：Html树结构特性、新闻网页标签的中心化，新闻文本标签的连续性，新闻文本标签的层次性，新闻文本标签的文本特征，Html修饰标签特性等。

（3）提出一种非新闻网页自识别方法。为了确保新闻采集过程中获取网页的内容都是新闻网页的文本，用于对网页进行判断，过滤非新闻网页。该方法的中心思想基于以下特点：非新闻网页正文长度偏短且均匀性、非新闻文本标签的零散性、非新闻文本标签的多中心性、非新闻标签多修饰标签性等。

（4）提出基于智能模版的新闻网页正文提取方法。该方法的主要思想是根据特定门户新闻网页正文的特点，通过机器学习的方法，自动适配网页正文提取的模版，再根据配置的模版对该门户类网页进行正文提取。

通过基于聚焦网络爬虫的智能Web新闻采集系统的设计与实现，及对新闻采集效果分析，证明了本文提出非新闻网页自识别方法和新闻网页文本提取方法的正确性和通用性，（也证明本文的爬虫策略在网页选择方面具有一定的效率），为智能新闻采集提供了新的方法和思路。

**关键词**：网络爬虫，非新闻网页自识别，新闻网页正文提取，标签特征

ABSTRACT

In recent years, as a research field of reverse engineering, point cloud space reconstruction technology has been greatly developed. It is largely used in many areas such as CAD modeling, geological modeling and so on. With the applications of surface reconstruction technique in real life are more widely, a large number of surface reconstruction algorithms for each different application areas have been brought up .In geologic surface reconstruction, the main problem faced is : firstly, there are a variety of complex constraints; secondly, for some large work area (such as the Sichuan Basin and other basins level work area), geological data is too large, ordinary PC memory is limited, it is difficult to achieve surface reconstruction using conventional method.

This thesis launches the research on large scale seismic data point cloud surface reconstruction problem under a variety of geological constraints. Implement a large-scale point cloud surface reconstruction method facing the complex constraints. The main contents of this article are in the following:

1. The thesis presents a surface reconstruction method for complex constraints. Based on point cloud surface reconstruction method on the geological boundary, the thesis made the research about the complex constraints appear in practical engineering like terrain reverse restriction and terrain trend restriction We made the research about constructing constraint of the regional geological regional, splicing reverse area and other areas, adding the terrain trend constraints, and finally achieved surface reconstruction.

2. The thesis presents a method of surface reconstruction for large scale seismic data. This thesis puts forward a method of using terrain block idea to solve problems. Divide massive earthquake data that can not be processed in single computer into the terrain block. Set the scheduling method of terrain blocks. Achieved an efficient and smooth process of surface reconstruction with mass seismic data. And made the deal effect.

目 录

[第一章 绪论 1](#_Toc413749731)

[1.1 研究背景及意义 1](#_Toc413749732)

[1.2 研究现状 1](#_Toc413749733)

1.2.1 网络爬虫研究现状

1.2.2 Web新闻内容提取现状＋新闻网页识别

[1.3本文的主要工作与贡献 4](#_Toc413749734)

[1.4论文章节安排 5](#_Toc413749735)

[第二章 网页正文提取及爬虫技术综述 6](#_Toc413749736)

2.1网页正文提取技术分析

[2.1.1基于DOM树的网页正文提取技术 6](#_Toc413749737)

[2.1.2基于视觉特征的网页正文提取技术 10](#_Toc413749740)

[2.1.3基于视觉特征的网页正文提取技术 17](#_Toc413749744)

[2.2 网络爬虫涉及的技术 20](#_Toc413749747)

[2.3本章小结 20](#_Toc413749747)

[第三章 基于文本标签特征聚类的网页正文提取方法（及新闻网页识别） 36](#_Toc413749758)

[3.1 引言介绍 36](#_Toc413749759)

[3.2 网页预处理及残损网页修复 36](#_Toc413749760)

3.2.1 HTML介绍

3.2.2 标签清理及修复

[3.3 标签特征及提取 42](#_Toc413749763)

[3.4 特征文本聚类及网页文本簇选择 44](#_Toc413749764)

[3.5 网页正文簇经验性调整 46](#_Toc413749765)

[3.7 新闻网页识别方法 48](#_Toc413749766)

[3.8 网页文本存储 48](#_Toc413749766)

[3.9 本章小结 48](#_Toc413749766)

[第四章 智能新闻采集策略设计 21](#_Toc413749748)

[4.0 初始种子选择策略 21](#_Toc413749749)

[4.1 初始种子选择策略 21](#_Toc413749749)

以文搜文，关键词搜索

[4.2 网页处理及正文提取 21](#_Toc413749750)

[4.3 新闻网页识别方法 23](#_Toc413749751)

[4.4 爬虫的搜索策略 25](#_Toc413749752)

[4.4.1 中文分词策略 26](#_Toc413749753)

[4.4.2 新闻主题相似度 28](#_Toc413749754)

[4.4.3 搜索策略（相似度越高则优先爬取） 31](#_Toc413749755)

[4.5 多线程抓取策略（线程池） 33](#_Toc413749756)

[4.6 本章小结 35](#_Toc413749757)

[第五章 智能web新闻采集实现及测试 50](#_Toc413749767)

[5.1 架构 50](#_Toc413749768)

[5.2 网页获取及解析模块 50](#_Toc413749769)

[5.3 实验及结果 51](#_Toc413749771)

[第六章 总结与展望 63](#_Toc413749773)

[6.1 工作总结 63](#_Toc413749774)

[6.2 工作展望 64](#_Toc413749775)

[致 谢 ............................................................65](#_Toc413749776)

[参考文献 66](#_Toc413749777)

[攻读硕士学位期间取得的研究成果 69](#_Toc413749778)

爬虫借鉴论文：面向医疗主题的智能网络爬虫的研究及实现

网页正文提取借鉴论文：1、基于视觉语义块的网页正文提取算法

2、基于标签分析的网页正文提取技术研究

# 绪论

## 研究背景及意义

随着Internet互联网技术的飞速发展，特别是Web应用的飞速发展，人们正面临着更具挑战的问题——“信息爆炸”，也即是信息极大丰富，信息极速传播，而知识过于贫乏，随着国家号召的“互联网＋”的号召，其将给人们带来更好的互联网体验，准确而迅速提取出网页的主题信息已经成为一个重要而有意义的研究方向。随着技术的发展，Web逐渐成长为内容生产与消费的平台，互联网上形成了无数以HTML网页形式存在的信息源。然而Web网页内容不仅包含用户关注的标题和正文，还包含大量与主题无关的噪声信息，如导航条、广告、推荐链接、免责声明和版权声明等噪声信息，使得用户不能快速准确的找到自己所需要的信息，这些噪声数据也大大妨碍了自动化程序对网页内容的检测、分析、挖掘和利用效果。近年来随着搜索引擎、网络舆情监控和智能Web等应用和研究工作的陆续展开，针对网页标题和正文等内容的抽取研究也受到广泛的关注，针对某一个新闻网页进行文本提取很简单，然而从成千上万种新闻网站中精准地抽取新闻网页内容将是一个极具挑战性的问题。

智能的Web新闻采集过程中，首先要从网络上获取网页源码，其次是对网页进行解析，获取新闻标题和正文部分内容。现有的爬虫技术已很完善，如何再进一步智能化，必然是从网页获取后的网页内容提取方面入手。虽然很多研究工作者在Web信息提取领域中已做了很多研究工作，人们对信息提取技术的研究已比较深入，但随着Web技术的发展和网站商业化运作越来越严重，噪声信息空前加大，使得很多方法已经不再适用于当今网页信息的提取，这也为网页信息的提取带来了新的挑战。

网络互联的信息时代，各门户网站、博客主页和论坛都是人们获取信息的主要途径，各类网页都有自己特有的网页格式，在新闻采集过程中，我们只关注包含有用信息的网页。非新闻网页包含网站引导页面、大部分BBS论坛页面、部分多图且无正文的网页页面或者多图有少部分对图片描述文字的页面和多数问答类型的页面等等，这类页面有共同的特征：网页文本信息几乎没有或很少、问答类页面的题目描述很短而回答很多、BBS论坛帖子的主题很短而回帖很多。在智能网页提取过程中这类网页的文本特征很弱，从而会造成高错误率，因此在新闻文本采集过程中进行非新闻网页识别的研究具有重要的意义。

## 研究现状

智能新闻采集简单的说就是根据需求自动化的从互联网上获取相关新闻网页，再智能的抽取网页中有用的信息，该抽取方法需要自适应各种各样的网站，避免人为提取。在采集过程主要涉及三个大方面的研究：

其一、网络爬虫。网络爬虫其实是一个基于HTTP协议的网络程序，这个名字的得来是因为它像爬虫一样在互联网这张以页面为节点、以链接为边的大网上“爬行"。爬虫从一些起始页面开始，在对这些起始页面进行分析以后，沿着从这些页面中找到的链接再对其它的页面进行访问。网络爬虫技术是搜索引擎之本，国内在爬虫方面的代表是百度，国外有强大的谷歌，实时的发现并爬取整个互联网实时变化的数据将决定搜索引擎的成败。网络爬虫根据其爬取方式可分为分布式和集中式爬虫，常见的爬虫策略有宽度优先、深度优先和聚焦最优优先的爬虫策略。

其二、Web新闻网页文本提取。（copy from：基于标签路径特征的Web新闻内容抽取研究－吴共庆）信息抽取(IE, Information Extraction)最早是作为自然语言处理的一个分支领域，始于20世纪60年代中期,研究从自然语言文本中获取结构化信息。它以两个长期的、研究性的自然语言处理项目为代表,一个是美国纽约大学开展的Linguistic String项目,另一个是耶鲁大学Roger Schank等开展的有关故事理解的研究。主流方法是采用自然语言处理技术从自然语言文档中定位特定的信息,利用文本中的语法和语义信息抽取出结构化数据。然而Web信息抽取与传统的文本信息抽取存在很多不同之处,因而,难以直接使用文本信息抽取技术进行Web信息抽取。目前为广大新闻采集工作者所熟知的网页信息解析技术是基于正则匹配的网页模版适配方法，该方法需要对每个待爬取的网站配置正文信息提取模版，当需要解析该网站的页面时根据模版提取对应的信息，但这种方法不具有通用性和实时适应性，当网站开发者做稍微改动，其模版就不可能失效，导致其运维成本很大；除了基于模版的方法外，还有基于DOM数解析的方法、基于视觉特征的方法和基于行块分布函数的方法来提取网页文本信息，各个方法都有其优缺点，后文会对其进行一一分析。信息抽取技术日渐成熟的同时,Web信息逐渐呈现爆炸式的增长,在海量Web信息中获取有用资源越来越困难,Web信息抽取逐渐成为研究的热点。然而,智能Web信息抽取与传统的文本信息抽取存在很多不同之处,如果在现有的技术上直接把智能文本信息抽取技术运用在新闻采集系统中，其抽取的新闻文本的可信度将产生很大的质疑。

其三、非新闻网页过滤，在新闻采集系统中，通过爬虫模块不停的向URL列表添加待爬去的网页URL，但是不是所有的URL对应的网页都是新闻网页，需要添加非新闻网页过滤模块。目前在新闻网页识别方面的研究主要从URL特征、网页结构特征、网页内容属性特征几个当面进行新闻网页识别，随着网站数量的突增和Html技术的发展，新闻网页识别需要结合当下的网页特点进行改进，以提高其识别效率。

## 1.3 本文的主要工作与贡献

本文首先系统性的介绍了智能Web新闻文本采集方法中所涉及的技术，再结合实际需求，确定新闻采集过程中的爬虫策略；其次回顾了目前已有的网页正文提取的主流技术，并对各种不同的算法进行分析比较，了解各方法的局限性和优缺点，然后在已有算法的总结基础上，提出一种新的新闻网页正文提取算法，与现有的算法进行比较，再而根据新闻采集过程中的需求衍生出新闻网页自识别的方法研究。

主要研究内容：

（1）在新闻采集系统中，采用聚焦爬虫策略，其语义分析过程中不简单依赖于关键字，而是通过对文本的各关键词的权重分析，通过计算文本相似性进而确定爬虫方向。

（2）提出基于文本标签特征挖掘的网页正文提取方法。该方法的中心思想基于以下特点：Html树结构特性、新闻网页标签的中心化，新闻文本标签的连续性，新闻文本标签的层次性，新闻文本标签的文本特征，Html修饰标签特性等。

（3）为了确保新闻采集过程中获取网页的内容都是新闻网页的文本，提出一种非新闻网页自识别方法，用于对网页进行判断，过滤非新闻网页。该方法的中心思想基于以下特点：非新闻网页正文长度偏短且均匀性、非新闻文本标签的零散性、非新闻文本标签的多中心性、非新闻标签多修饰标签性等。

## 1.4 论文章节安排

本文利用六章的篇幅对面向复杂约束的大规模点云曲面重建方法的研究与实现进行了详细的介绍：

第一章 绪论。对本文中涉及到的空间点云曲面重建技术的应用背景、研究意义和研究现状加以介绍，并简单介绍了本文的主要工作和论文框架。

第二章 地质曲面重建基础。介绍了地质勘探数据的特点，由此引出本文实现方法的难点所在。其后，对复杂约束下曲面重建中使用的约束Delaunay三角剖分技术进行了介绍。最后对大规模地震数据下的三维地质建模技术加以介绍。

第三章 面向复杂约束的地质曲面重建方法。此章详细介绍了倒转地质约束和地势走向约束的情况，主要考虑在倒转约束和地势走向约束下的点云曲面拓扑形成和曲面重建。

第四章 基于地形分块的大规模盆地级别地质曲面重建。讲述如何在盆地级别的大规模数据下，利用单台PC机完成曲面重建工作。

第五章 大规模地震点云曲面重建系统实现。该章讲述了本文提出的面向复杂约束的大规模点云曲面重建系统的算法子模块、系统界面与系统效果图、系统工作效率测试。

第六章 总结与展望。该章对本文提出的创新点和测试结论进行了系统性总结。对未来可能的研究方向进行了规划。

# 智能Web新闻采集技术基础

智能Web新闻采集是一种根据采集者的意愿进行新闻采集，对大千新闻网站的Web网页正文信息抽取过程中不用采集者做任何前期配置工作，而是需要一种自适应各种各样网页的文本抽取技术。智能Web新闻采集方法主要包含三方面内容：第一个是如何获取网页、根据什么策略进行获取网页，即网络爬虫技术；第二个是在获取新闻网页源码后如何对各式各样的网页进行信息抽取，即网页信息提取技术；第三个是如何识别哪些网页是非新闻网页，即非新闻网页识别技术。

## 网络爬虫相关技术

（copy from：<http://www.hyqb.sh.cn/publish/portal2/tab227/info4124.htm>）

（copy：面向医疗主题的智能网页爬虫的研究与实现）

1994年世界上第一个网络检索工具Web Crawler问世，目前比较主流的搜索引擎有百度、Google、Yahoo、Info seek、Inktomi、Teoma、Live Search 等。出于商业机密的考虑，现在各个搜索引擎使用的Crawler 系统的技术内幕一般都不公开，现有的文献资料也仅限于概要性介绍。随着网络信息资源呈指数级增长及网络信息资源动态变化，传统的搜索引擎提供的信息检索服务已无法满足人们日益增长的对个性化服务的需求，正面临着巨大的挑战。以何种策略访问网络，提高搜索效率，已成为近年来专业搜索引擎网络爬虫研究的主要问题之一。

### 网页抓取策略

网页抓取策略类似于图的遍历方式，根本便利的方式不同可以分为三种，即深度优先策略、聚焦策略（最近优先策略）、宽度优先策略。

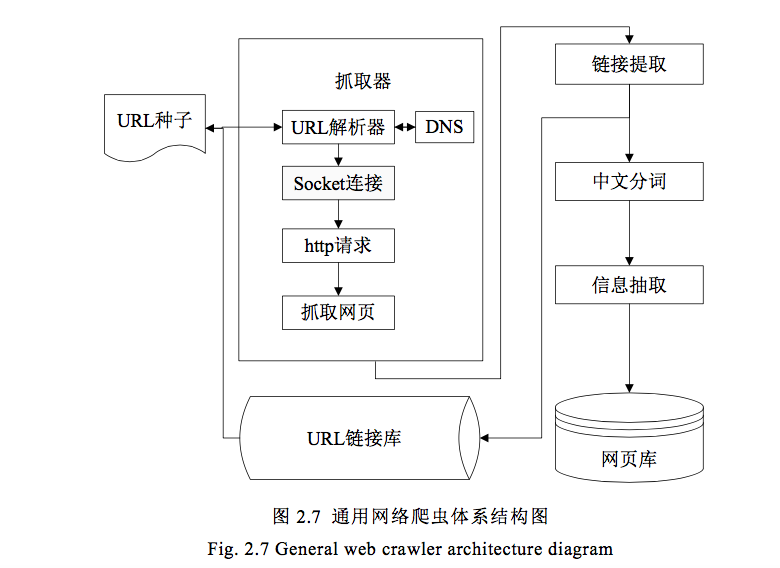
（一）深度优先策略。深度优先策略是在开发爬虫早期使用较多的方法，它的目的是要达到被搜索结构的叶结点，即那些不包含任何超级链接的HTML文件。在一个HTML文件中，当一个超级链接被选择后，被链接的HTML文件将执行深度优先搜索，即在搜索其余的超级链接结果之前必须先完整地搜索单独的一条链。深度优先搜索沿着HTML文件上的超级链接走到不能再深入为止，然后返回到某一个HTML文件，再继续选择该HTML文件中的其他超级链接。当不再有其他超级链接可选择时，说明搜索已经结束。其优点是能遍历一个Web站点或深层嵌套的文档集合。缺点是因为Web结构相当深，有可能造成一旦进去再也出不来的情况发生。

（二）聚焦策略；聚焦爬虫的爬行策略只跳出某个特定主题的页面，根据“最好优先原则”进行访问，快速、有效地获得更多的与主题相关的页面，主要通过内容与Web的链接结构指导进一步的页面抓取。聚焦爬虫会给它所下载的页面一个评价分，根据得分排序插入到一个队列中。最好的下一个搜索对弹出队列中的第一个页面进行分析后执行，这种策略保证爬虫能优先跟踪那些最有可能链接到目标页面的页面。决定网络爬虫搜索策略的关键是如何评价链接价值，即链接价值的计算方法，不同的价值评价方法计算出的链接的价值不同，表现出的链接的“重要程度”也不同，从而决定了不同的搜索策略。由于链接包含于页面之中，而通常具有较高价值的页面包含的链接也具有较高价值，因而对链接价值的评价有时也转换为对页面价值的评价。这种策略通常运用在专业搜索引擎中，因为这种搜索引擎只关心某一特定主题的页面。

（三）宽度优先策略；在宽度优先搜索中，先搜索完一个Web页面中所有的超级链接，然后再继续搜索下一层，直到底层为止。例如，一个HTML 文件中有3个超级链接，选择其中之一，处理相应的HTML文件，然后不再选择第二个HTML文件中的任何超级链接，而是返回，选择第二个超级链接，处理相应的HTML文件，再返回，选择第三个超级链接，并处理相应的HTML文件。一旦一层上的所有超级链接都被选择过，就可以开始在刚才处理过的HIML文件中搜索其余的超级链接。这就保证了对浅层的首先处理。当遇到一个无穷尽的深层分支时，不会导致陷进WWW的深层文档中出不来的情况发生。宽度优先搜索策略还有一个优点，它能在两个HTML文件之间找到最短路径。宽度优先搜索策略通常是实现爬虫的最佳策略，因为它容易实现，而且具备大多数期望的功能。但是如果要遍历一个指定的站点或者深层嵌套的HTML文件集，用宽度优先搜索策略则需要花费较长时间才能到达深层的HTML文件。

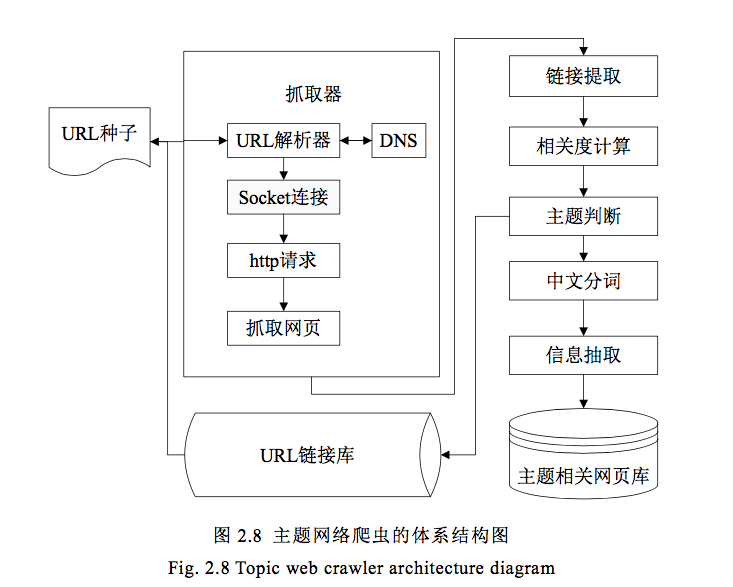
## 网络爬虫分类

按照爬虫的功能不同、爬行策略、实现技术和结构差异的不同，网络爬虫可以划分为三大类:通用爬虫、聚焦爬虫（主题爬虫）、深度爬虫。

（一）通用爬虫采用宽度优先策略，从给定的初始网页开始，抓取初始网页源码，从网页源码中提取出新的URL，将新的URL不进行任何排序直接加入待爬取网页URL队列中继续爬取，一直向整个互联网扩展，直到满足预设的终止条件为止。通用爬虫采用的策略对URL没有进行任何过滤，导致抓取的网页很多都是无用的，从而使爬虫的效率异常的低。

（二）聚焦爬虫又称为主题爬虫，从网页中取出新的URL后不会直接放入到待爬去URL队列中，而是经过设定的网页分析算法对链接进行分析，过滤掉于预设网页主题不符合不相干的网页，选择与预设网页主题相关度高的网页URL继续进行爬取过程。聚焦爬虫使用的是最佳优先策略。 由于对网页URL进行了过滤,下载的网页都是针对预设主题相关的网页,可以避免下载无用网页带来的网络消耗,减少对无用网页的存储,使爬行的效率大大提高。

聚焦爬虫的性能取决于网页过滤策略的效果，好的网页过滤策略将很好的过滤无用网页和与主题不相关的网页，反之，则过滤掉与主题相关的网页或过滤后的网页抓取后仍然是无用的网。现有的网页过滤策略分为基于超链接评价的策略，如PageRank算法和HITS算法等；基于网页内容评价的策略，如Fish search算法和Shark search算法。



（三）深度爬虫（删去？）

网络爬虫最新发展依附于Web2.0的发展。传统的网络爬虫技术主要应用于抓取静态Web 网页，随着AJAX/Web2.0的流行，如何抓取AJAX 等动态页面成了搜索引擎急需解决的问题，因为AJAX颠覆了传统的纯HTTP 请求/响应协议机制，如果搜索引擎依旧采用“爬”的机制，是无法抓取到AJAX 页面的有效数据的。AJAX 采用了JavaScript 驱动的异步请求/响应机制，以往的爬虫们缺乏JavaScript语义上的理解，基本上无法模拟触发JavaScript的异步调用并解析返回的异步回调逻辑和内容。另外，在AJAX的应用中，JavaScript 会对DOM结构进行大量变动，甚至页面所有内容都通过JavaScript 直接从服务器端读取并动态绘制出来。这对习惯了DOM 结构相对不变的静态页面简直是无法理解的。由此可以看出，以往的爬虫是基于协议驱动的，而对于AJAX 这样的技术，所需要的爬虫引擎必须是基于事件驱动的。

## 2.2 Web新闻文本提取技术

（copy from：基于标签路径特征的Web新闻内容抽取研究－吴共庆）

在网页信息抽取技术日趋成熟的同时，Web网页信息却呈现出爆炸式的增长，在数以亿计的Web网页信息中针对性的获取有用资源变的越来越困难，当下Web网页信息抽取逐渐成为众多工作者的研究的热点。目前主流的Web网页正文提取技术大致有以下几 种方法：基于包装器的方法、基于DOM树的方法、基于视觉特征的方法、基于文本特征的方法和基于行块分布函数的方法。

## 2.2.1基于包装器的方法

copy from：基于多特征融合的网页正文提取及双语网站探测

这类方法特征为针对不同的网页源生成其网页模板,进而利用网页模板和规 则得到网页的包装器(wrapper),然后利用包装器对数据源进行解析。代表成果主 要有以下几个:

1997 年美国斯坦福大学的 J.Hammer,M.Brenning 等人利用基于模板的包装器 (wrapper)处理和挖掘 TSIMMIS( Stanford 研究开发的信息系统,其采用异构方 式存储数据源)系统内信息[18],为其后的研究学者利用包装器进行网页信息分析 提供了方向,包装器是个特殊的程序,程序内部包含规则,而规则符合一定的模 板样式,利用这些规则模板去海量信息源中抽取与有用的信息,包装器内部的规 则主要有三个来源:一是通过人工抽取;通过需要人为指导的半自动方法进行抽 取;通过全自动方法抽取。此论文中的 J.Hammer,M.Brenning 等人通过人工输入 方式获得规则,利用三元组[结果,输入,模板]方式从数据库系统中抽取信息[18]。

1997 年美国南加州大学的 Craig A 和 Naveen Ashish 利用中介程序来从特定 领域的多个网站数据源抽取有效的信息[16]。在多个网站数据源中每个数据源都包 含自己的包装器和抽取方法,中介程序把一个查询针对不同的数据源处理成符合 相应包装器的查询语句,最后将来自各个数据源的结果整合获得最终结果。同时 利用结果自动扩展不同数据源的包装器规则。此算法主要包含三个步骤:

a) 结构化数据源 针对不同数据源采用人工方式标注少量有价值网页信息 块和字块,然后利用程序自动标注。

b) 建立解析器(Parser)在结构化数据源后为的数据源建立各自的分析程序。

c) 和中介程序建立连接 为各个数据源生成解析器后中介程序作为整合各个 数据源的中介必须能够和各个数据源的包装器建立连接,同时能够调用不 同数据源内部的数据抽取程序。

2003 年日本九州大学的 Daisuke Ikeda 和 Yasuhiro Yamada 等人基于 HTML网页结构的特殊性文献完成一个全自动的包装器生成工具 RoadRunner[19],此工具把 HTMl 代码中的内容泛化成字符串、关键字、链接、段落、循环等结构进而生成网 页包装器,此法不用用户输入任何知道信息之需要几个初始网页训练便可得到网页模板。

采用包装器的的提取方法实质是采用规则和模式匹配的方法进行抽取,针对单 一格式网页和数据效果很好,但是由于 HTML 语言灵活多变,规范出现较晚且影 响较小,基于包装器的方法很难做到通用,同时数据源的更新会造成包装器程序 维护耗费大量的人力物力。

## 2.2.2基于DOM树的方法

（CP基于视觉语义块的网页正文提取技术算法研究）

这类技术主要是从网页的HTML源文件代码角度来对网页进行分析,首先根据网页的HTML标签特征以及嵌套关系,将网页的各个元素构建成一棵DOM树, 其次根据各种规则对该DOM树进行剪枝筛选,其中包括了中文与超链接比例分析、链路分析、<table>块分析以及结点统计分析等各种方法,计算树的每个子树的正文概率，再次确定了网页正文信息所在的区域,然后提取出正文信息。 典型的代表成果有以下几个：

周佳颖等人于2009年提出了基于统计与正文特征的中文网页正文抽取方法【6】。该方法首先基于统计的方法来得到一条正文路径，然后通过对该路径的正文特征进行研究，识别网页的正文区域以及子树主干，最后再根据正文区域以及该子树主干所具有的正文特征来获得完整的正文信息。

安增文等人于2010年提出了基于机器学习的网页正文提取算法【9】。该算法在将网页解析成DOM树之后，对DOM树种每一行文本的文本密度以及文字内容与标题的相关度等值进行计算，然后将计算结果作为输入参数传入BP神经网络，进而提取网页正文信息。

张瑞雪等人于2011年提出了逆序解析DOM树及网页正文信息提取算法【13】。该算法按照倒序从后向前对网页信息进行解析，构建DOM树，然后按照正文信息在网页中的分布特征确定部分正文信息在DOM树中的位置，然后根据树结构的相似性寻找出其他正文信息所在位置，最后将之整合，提取出网页正文内容。

王超等人于2012年提出了基于CURE算法的网页分块及正文块提取研究【14】。该算法在将网页解析成DOM树之后，对DOM树的结点添加新的属性，使之转化成为带有信息结点偏移量的扩展DOM树，然后利用CURE算法对这些结点进行聚类，每个聚类的结果簇就代表着页面的不同块，然后根据正文块的三个特征：含有较少的超链接数；含有较多的长信息结点；含有较多的数据记录条数，来构造信息块的权值公式，最后再利用该公式识别出正文块，从而提取出网络正文信息。

## 2.2.3基于视觉特征的方法

（CP基于视觉语义块的网页正文提取技术算法研究）

与基于DOM树的正文提取技术不同,基于视觉的正文提取技术主 要是从用户视觉的角度出发,对网页结构进行解析,进而提取出网页正文信息。 这类技术虽然也需要创建DOM树,但是其构建树的主要依据并不是HTML标签 信息,而是根据用户浏览网页的体验,按照语义块信息进行构建。 其代表成果有如下几个：

微软亚洲研究院于2003年提出的VIPS(Vision-based Page Segmentation) 网页分割算法。VIPS算法主要是针对之前主流方法创建的DOM树无法完全体现网页页面的 语义结构而提出的,该算法主要是从用户视觉的角度出发,不考虑具体的HTML 源码信息,而是根据用户感官体验来将网页进行划分,通过这种方式,该算法可 以从网页的DOM树中提取出一个个独立的语义块,然后再根据这些语义块的具 体大小以及坐标信息,检测出存在于这些语义块之间的分隔条,包括了水平分隔 条与垂直分隔条,然后再根据这些分隔条的信息,对页面的语义结构进行重构。 对于重构结果中的每个语义块,又可以再次采用之前的流程对其进行更细致的分 割,生成更小的语义块,如此反复直至分割精度达到预定要求,然后就可以根据 页面分割的结果,提取出所需的正文信息。

黄玲等人于2008年提出了基于网页分块的正文信息提取方法【l 8】。该方法首先提取出网页标题，然后根据网页正文与主题文本标题匹配的方式来对网页的正文内容块进行识别和提取，然后再利用正则表达式以及一些简单的判别规则来过滤掉内容块中的HTML标签和一些无关文字。由于该方法比较注重网页标题与正文的相关度，因此对于一些没有准确定位网页正文标题的网页，则可能会发生内容块提取错误，进而导致正文提取失败。

安增文等人于2010年提出了基于视觉特征的网页正文提取方法[19]。该算法主要是基于了VIPS算法，针对VIPS算法需要处理大量语义块而导致的性能问题，该算法采用了逐层分块逐层删减的方法，因而减少了对那些已经确定与正文内容无关的噪音语义块的处理。同时，为了减少工作量，降低复杂度，该算法也并没有在算法循环中进行分隔条的检测，以及网页语义结构的重构，而是将这些工作都保留到了正文过滤的结束阶段，最后再对提取出来的正文语义块进行相应的处理。

## 2.2.4基于行块分布函数的方法

copy from：基于多特征融合的网页正文提取及双语网站探测

哈尔滨工业大学的陈鑫第一次提出这个方法,这方法应该是目前所有正文提 取算法中最快的,因为这个算法的复杂度是 O(n)的。此算法有两个重要的依据: 正文区的密度、行块的长度。此算法对网页的处理主要包括:取出网页的标签, 把余下的正文当作 Ctext;在 Ctext 中的行号为中心,获取周围的几行作为一个行 块(Cblock);计算行块分布函数,其中自变量 x 为行的编号,因变量 y 为以 x 为 轴的此行块中的 Ctext 长度。根据行块分布函数寻找骤升点和骤降点。骤升点的斜 率要大于一个特定的阈值 p,骤降点的斜率要小于一个特定阈值 q。取骤升点和骤 降点之间的 Ctext 作为网页正文。 此算法的最大特点就是速度快,因为时间复杂度为 O(n),作者在 1000 篇主题 网页中进行测试,其中平均网页长度大于 1000 行得到时间消耗为 21.29S。准确率在95%以上。

## 2.2.5基于文本特征的方法

（CP基于视觉语义块的网页正文提取技术算法研究）

与之前构建DOM树或视觉DOM树的技术不同,这类技术通常只需要对网页HTML源文件进行分析,然后根据网页的各种正文文本特征,创建各 种过滤规则来对网页进行过滤,进而提取出网页正文信息。由于减少了DOM树 的创建分析遍历等过程,因而这类方法通常在时间性能上较为出色,然而由于对网页的正文文本特征有一定的要求,因而其健壮性稍有欠缺。 其代表成功有如下几种：

常红要等人于2010年提出了基于HTML标记用途分析的网页正文提取算法 【25】。该算法主要分为以下三步:一、删除网页HTML源文件<head>区域块内 与网页标题无关的内容;二、通过获取到的网页标题与<body>区域块中的网页文 本进行匹配。确定<body>区域块中网页正文标题所在位置:三、根据标题所在位 置与文本匹配规则,删除<body>区域块中与网页正文无关的内容,最后就能顺利得到网页的正文信息。

王利等人于2010年提出了基于内容相似度的网页正文提取算法【261。该算法主要是利用了文本相似度的计算方法,其中心思想为网页正文文本与标题文本的 内容相似性,通过计算网页HTML各标签结点中的文本内容与其各级标题的相似 度,来判断文本信息的有用性,从而进行网页的筛选与正文的抽取,获取网页的 文本信息。该算法相当依赖于网页标题和网页中大文本块的正确抽取,因而对该阶段的算法性能效率有较大要求。

欧阳佳等人于2011年提出了基于DBSCAN算法的网页正文提取算法【28】。该 算法主要是基于分段因子对网页源码进行过滤获取纯文本段,然后将每一段都视 为二维空间中一个点,通过采用DBSCAN聚类算法来对这些点进行聚类分析获 取正文内容。该算法对于网页正文较为紧凑,字数分布较为均匀的网页有着很好 的效果,但是对于一些字数很少的正文段则可能会发生遗漏问题。

## 2.3非新闻网页识别技术

过往非新闻网页识别研究主要基于URL特征、网页机构特征和内容属性特征进行识别。基于URL特征识别新闻网页，其首先对新闻网页的URL地址特征进行建模，运用机器学习的方法自动提取新闻网页的URL特征构造新闻网页URL特征库，再通过新闻网页URL特征库和新闻网页URL特征识别机制判断给定的URL是否是新闻网页的URL，此方法效果的好坏在于对URL特征工程的挖掘，目前互联网网站数量的突增，各个网站的URL地址设计各不相同，很难挖掘出一个通用的URL特征库，造成网页URL识别的可信度降低。

对于一部分非新闻网页的判断，可以通过URL特征进行预判断。静态网页的URL一般是以 .htm 、.html、.shtml、.xml等形式为后缀结尾，大部分非新闻网页都不是静态页面，通过判断URL是不是静态网页可以对非新闻网页进行一层过滤，但是对于一句话新闻网页或多图无正文网页来说，将没有合适的办法进行过滤。本文根据这类非新闻网页文本标签的特点，并对其特征进行相应挖掘，寻找特点规律进行非新闻网页的判断。

## 2.4 本章小节

本章讲述了智能Web新闻采集过程中所涉及到的网络爬虫和Web 新闻抽取技术。首先介绍了网络爬虫技术，包括爬虫策略和常见爬虫分类，然后介绍了现有的Web新闻抽取技术，介绍了各种技术对应方法的优缺点，为本文的文本提取方法的提出做铺垫。

# 基于文本标签特征挖掘的网页正文提取方法

网页文本提取技术在搜索引擎方面、提高新闻网页过滤与总结、情报监控与分析等Web应用系统有着十分重要的作用，而一个优秀的网页正文提取方法能够提高网络搜索的准确度和效率值，有效地提升用户体验值。然而,现有的网页正文提取方法或多或少存在一定的弊端、局限性或者不具有普适性,都还存在一定改进的空间。因此,通过总结研究现状特点，本章提出了一种新颖的新闻Web信息提取方法——基于文本标签特征挖掘的网页文本提取方法。在此基础上，提取一种新闻网页自识别方法，解决了采集过程采集到非新闻网页文本的弊端，使得新闻采集结果具有高可信度。通过实验验证，随着网页数量和网站数量的增加，基于文本标签特征挖掘的网页正文提取方法具有很强的通用性，其新闻正文提取正确率可以达到98%以上，再结合新闻网页自识别算法，总体正确率可以达到99%，由此可见该算法适用于智能Web新闻采集系统。

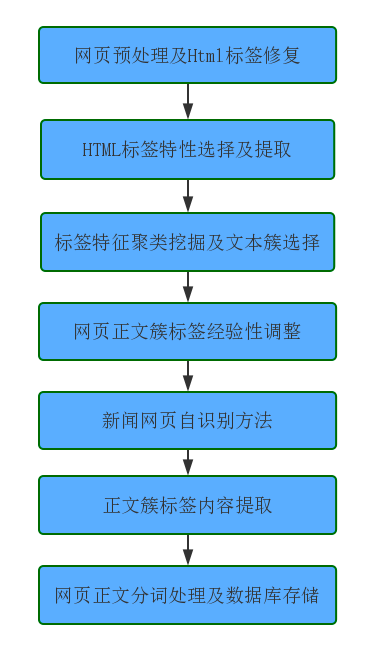
本章其余部分组织如下：第2.2节描述了网页预处理和残损网页修复过程中所涉及到的背景知识和相应的修复策略；第2.3节介绍了网页文本标提取及其特征挖掘过程；第2.4借描述了文本特征挖掘所涉及的算法和网页正文簇的选择策略；第2.5节介绍了在确定网页正文簇后的正文标签名及层次的确定，又介绍根据获取的正文标签标签进行经验性调整策略；第2.6节描述了新闻网页自识别的由来及识别方法的详细过程；第2.7节介绍网页文本提取后的存储及为语义挖掘做进一步分词处理；第2.8节时本章的小结部分，概括本章的大意。

## 方法思想及流程

大千网站几乎清一色的都使用统一的HTML协议，其新闻网页的布局具有规律可循。从搜索引擎的角度出发，目前互联网上网页大致分为以下两种：一、以超链接跳转为主的目录型网页，二、以网页正文文本信息为主的主题型网页，也称为新闻网页。

第一类是以引导跳转为主的目录型网页，其通常都是网站内部信息搜索页面、网站首页、网站的主题栏目信息页面等页面，这类网页的主要特点是整个网页布满了跳转标签，大量的超链接信息可以链接到其他相关的网站页面，网民通过额外的点击才能真正达到自己想要的浏览的页面，这类网页其内容的提取毫无意义，本文可以通过新闻网页自识别方法进行过滤。

第二类是以文本信息为主的主题型网页，其主要信息一般都是直接了当的展示在用户浏览界面上，用户点开页面即可获取想要的内容信息，该类网页的文本信息标签具有一定的规律，其特征与其他噪声标签有着明显的区别，我们可以从这个角度作为出发点进行特征挖掘。经过大量的实例分析和进一步研究发现，新闻网页中的各Html标签具有可分类性，进一步挖掘新闻网页文本标签特征后，发现正文标签具有连续性、中心化特性和层次性，其文本内容长度相比于噪声文本长度要长，且具有分句、段落性质，为此提出了一种基于文本标签特征挖掘的网页正文提取方法，同时也衍生出新闻网页自识别方法，用于噪声网页的自动户过滤。



算法流程图

文本标签特征挖掘方法相对于其他新闻网页文本抽取方法而言，具有更好的通用性，具有更高准确率，具有易用性且与网页语言无关性；在现实数据集上的实验结果验证表明，基于文本标签特征挖掘的网页文本提取方法能够满足新闻网页内容抽取的实际应用的需求。

## 原始HTML网页预处理及修复

## 2.2.1 HTML语言标准及特点分析

超级文本标记语言（HTML，HyperText Markup Language）是为“网页创建和其它可在网页浏览器中看到的信息”设计的一种标记语言,是标准通用标记语言下的一个应用，也是一种规范，一种标准，它通过标记符号来标记要显示的网页中的各个部分。HTML做为当今网页最常用的标记语言，其包含了网页的几乎所有的信息，可以很好的描述网页文档信息。在HTML语言中，有各种各样的标签名，例如<head>，<body>，<table>，<title>，<meta>，<script>，<div>等标签，这些有些标签是用来修饰文本格式，有些标签是用来存储JS(javascript)语言信息，有些是网页文本信息标签，如此功能各异的标签组成了HTML语言中最基础也是最重要的部分。比较常见的结构性标签包括如下几个：

html元素：标记HTML内容的开始和结束。

head元素：标记HTML文件头，包含不在正文中显示的关键字、标题、脚本等等。

title元素：标记HTML文件的标题。

body元素：标记HTML文件正文的开始和结束。

**以下表格列出了 HTML head 元素：[3]**

|  |  |
| --- | --- |
| **标签** | **描述** |
| <head> | 定义了文档的信息 |
| <title> | 定义了文档的标题 |
| <base> | 定义了页面链接标签的默认链接地址 |
| <link> | 定义了一个文档和外部资源之间的关系 |
| <meta> | 定义了HTML文档中的元数据 |
| <script> | 定义了客户端的脚本文件 |
| <style> | 定义了HTML文档的样式文件 |

其中<body></body>网页中显示的实际内容均包含在这2个正文标记符之间，正文标记符又称为实体标记，如<p>文本内容</p>,<h>标题文本信息</h>等。

HTML语言具有树结构特点，其每个标签都可做为一个树节点，标签具有层次的特点，Html网页源码中的新闻网页的正文标签一般处于同一个层次，而且新闻网页的正文标签大体是连续的，不排斥正文标签之间嵌套一些修饰文本格式个标签，例如a标签、td标签和li标签等。

## 2.2.2 标签预处理

本文的网页文本提取方法是围绕网页文本标签特征展开，而网页标签中包含大量无用的噪声标签，在后文中的特征学习过程中，这些无用的噪声标签会影响正文标签的学习过程，故在提取标签特征前需要排除用于Javascript语言的script标签、排除用于结构特征的style标签、排除noScript标签、排除标签注释内容、排除无用的表格span标签及内部列表li标签、排除文本格式修饰标签及换行标签等噪声标签。

从前人研究结果发现，基于Dom树解析的方式可以处理无用标签，但是在讲网页HTML源码转为Dom树的过程耗时比较大，会影响网页处理的效率，故本文中采用字符串处理的方式来进行网页标签的预处理工作，在保证效率的同时，质量上也很有保障。

在标签预处理过程中，删除标签要考虑两种情况：一、标签及标签之间的内容都要清理；二、只清理标签信息，标签之间的内容不清理。第一种标签及标签之间的内容都要清理的标签包括，html、script、noscript、style、注释等标签来；第二种只清理标签信息，标签之间的内容不清理，这类标签包括，span、strong、br、nbsp、em、b等标签，这类标签主要用于修饰段落、治表和文本字体信息，其内容可以规划到上一个标签中。

基于Dom树解析删除特定标签的方法进行标签预处理，可以运用现有的库进行操作，但是很耗时，故本文采用字符串解析的方式进行标签删除。

由于本方法采用的字符串处理方法进行标签清理，不能用Dom树解析库进行相应标签的清理，在此要声明两个定义：一、左标签，HTML标签的开始部分，例如<script>；二、右标签，HTML标签的闭合部分，例如</script>。根据标签处理的需求，本文定义了标签结构体如下：

struct lableMsg {

string lableName; //标签名

int beginIndex; //标签开始位置

int endIndex; //标签结束位置

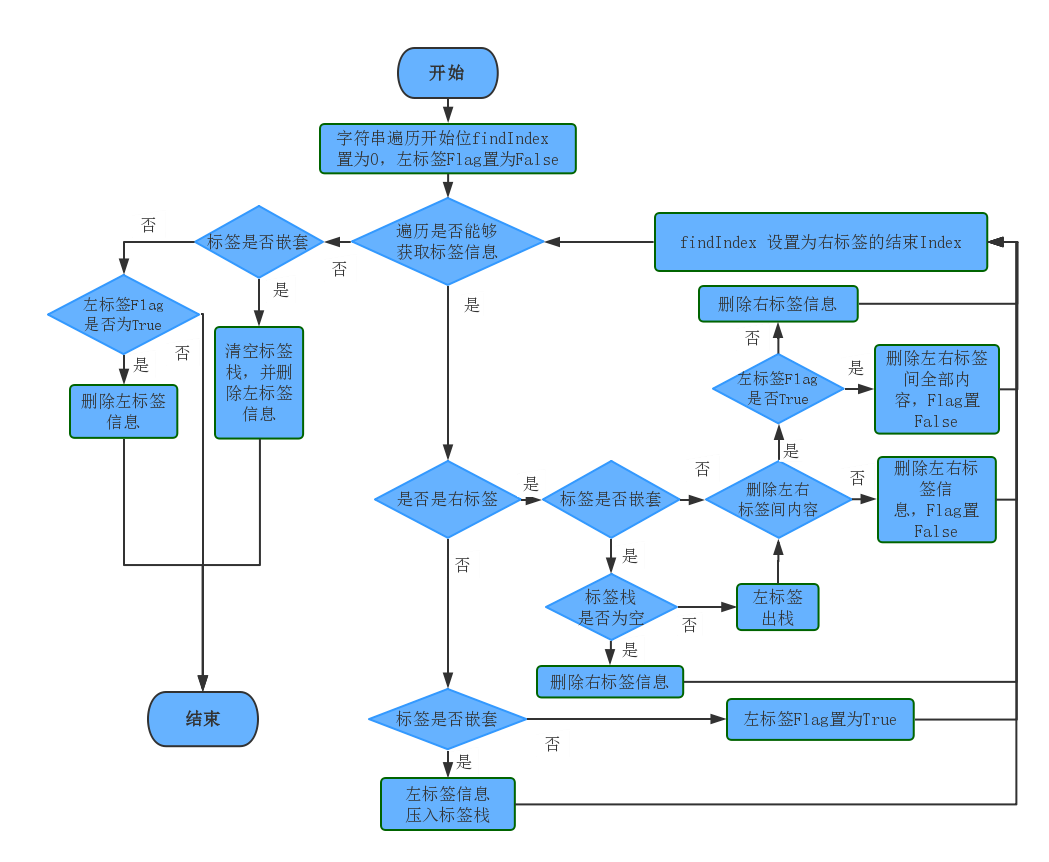
int lableFlag; //左右标签标识

bool leafLableFlag; //是否是叶子标签或是否被左左标签提取过

};

在遍历网页字符串获取标签时，根据HTML语言标签的特点，自动确定是做标签还是右标签，通过leafLableFlag字段去确认；为了在清理标签时快速的定位标签位置，在结构题中加入了标签的开始坐标和结束坐标。

根据不同的HTML标签的特点不同，清理噪声标签时可以分为两种情况：一、待清理HTML标签不具有嵌套性质，可以直接找到该标签的开始标签和结束标签，根据上述删除标签的两种情况进行标签清理；二、待清理标签具有嵌套性质，即有如下<p><p></p></p>标签嵌套情况，针对这种情况本文将借用栈的思想进行删除，当获取标签的左半部分时，直接将左标签信息压入栈内；当获取标签的右半部分时，考虑标签栈是否为空，如果为空，则删除右标签的标签信息，如果不为空，则将栈顶标签元素出栈，将栈顶的左标签和刚遍历到的右标签进行配对，根据上述删除标签的两种情况进行处理；循环遍历网页字符串，直到不再存在待删除的标签信息。标签处理的方法流程图如下：



标签处理的方法流程图

由于HTML标签对大小写不敏感，各网站的开发人员有自己的编码习惯，其HTML标签的大小写无规律可循，故在标签清理之前，将网页源码中的大写标签统一改为小写标签，如<SCRIPT></SCRIPT>改为<script></script>。通过对比后文算法的聚类效果，标签的过度清理或者不清理标签都将影响文本提取的结果，本文通过大量对比实验得出预清理的标签有以下：解释标签、head标签、script标签、style标签、noscript标签、span标签、strong标签、br标签、nbsp标签、font标签、em标签、img标签和b标签等。

提取的标签信息截图

## 3.2.3 标签修复及转义字符处理

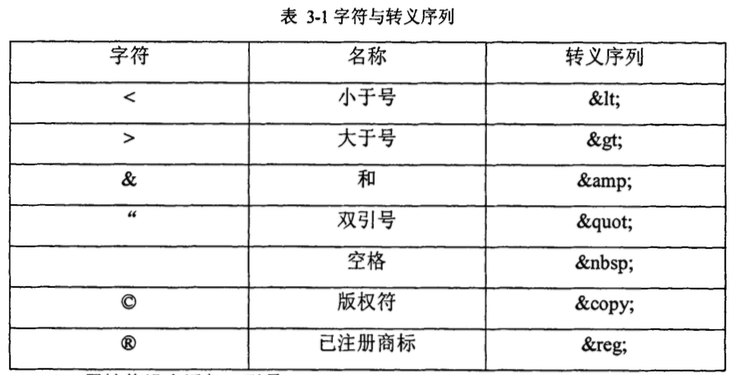
由于HTML语言标准规范对其语法的要求不是很苛刻，这就滋生了开发过程中书写HTML语句时的不严谨性，进一步导致网页源码中部分源代码上会出现不规范性，这种错误对浏览器而言无关紧要，因为浏览器具有一定的容错能力，能够将语法错误的语句解读并按照网页开发者的意愿显示在页面上，在上文提到的文本标签预处理过程中也可能照常部分标签不对称，这些错误对本文下文将提到的文本标签特征提取部分的工作而言，将对其正确性和和效率造成一定的影响。目前常见的HTML源代码不规范性主要包好以下几个部分：

（1）标签缺失

通常HTML的标签对有两种闭合方式，一种是标签成对出现，并在后一个标签开始处的“<”之后添加“/”符号将其闭合；另一种则是单独出现，需要在标签结束前的“>”之前添加“/”符号将其逼合。所谓的标签缺失，在第一类标签体现为只有左标签信息而缺失右标签信息，或则是挚友右标签信息而缺失左标签信息；第二类标签体现为“>”之前缺失“/”符号。

（2）特殊字符的转义

在HTML规范中“<”,“&”,“>“等字符都具有各自特殊含义，前两个字符用于标签,最后一个字符用于字符转义,这类字符并不能直接被用于内容编写使用,因此在使用这部分特殊字符时,需要将它们进行转化为转义序列，其转义如下表所示：



针对上文第一种情况而言，为了保证下文标签特征提取的完备性，在标签预处理之后需要对网页进行标签对缺失修复和特殊字符进行转义。标签对修复规则如下：由于标签具有嵌套的特点，本文在修复标签对时运用栈进行标签存储，从预处理后的网页字符串的第一个字符开始，依次取字符串中的左右标签，当获取左标签时，直接将标签信息压入栈顶；获取到右标签时，若栈为空，则在文本字符串中删除该右标签信息字符串，反之则取出栈顶标签，如果栈顶元素的标签名和当前获取的右标签名相同，则将栈顶元素出栈，如果栈顶元素的标签名和当前获取的右标签名不相同，则在预处理后字符串中删除右标签信息，因为这个标签缺失左标签，删除该右标签不影响下文的算法效果；当预处理后的网页字符串遍历完成后，若标签栈为空，则此时的预处理后的网页字符串中的标签对已经完备，若不为空，则从栈顶依次取出标签信息，在预处理后的网页字符串尾部添加该标签名对应的右标签，例如添加“</lableName>”字符串。

针对上文第二种情况而言，由于网页中部分文本字符是经过网页转码的，如果不进行相应的反转义，后文中提取到的网页文本将出现转义序列，为了避免因为转义字符造成提取到的网页文本信息不一致性，故需要对特殊的转义字符进行反转义，将转义序列转化为转义字符。本文是通过字符串处理的方式进行字符反转义的，不停在网页字符串中寻找转义序列，并将其替换为相应的转义字符。

## 文本标签特征选择及提取

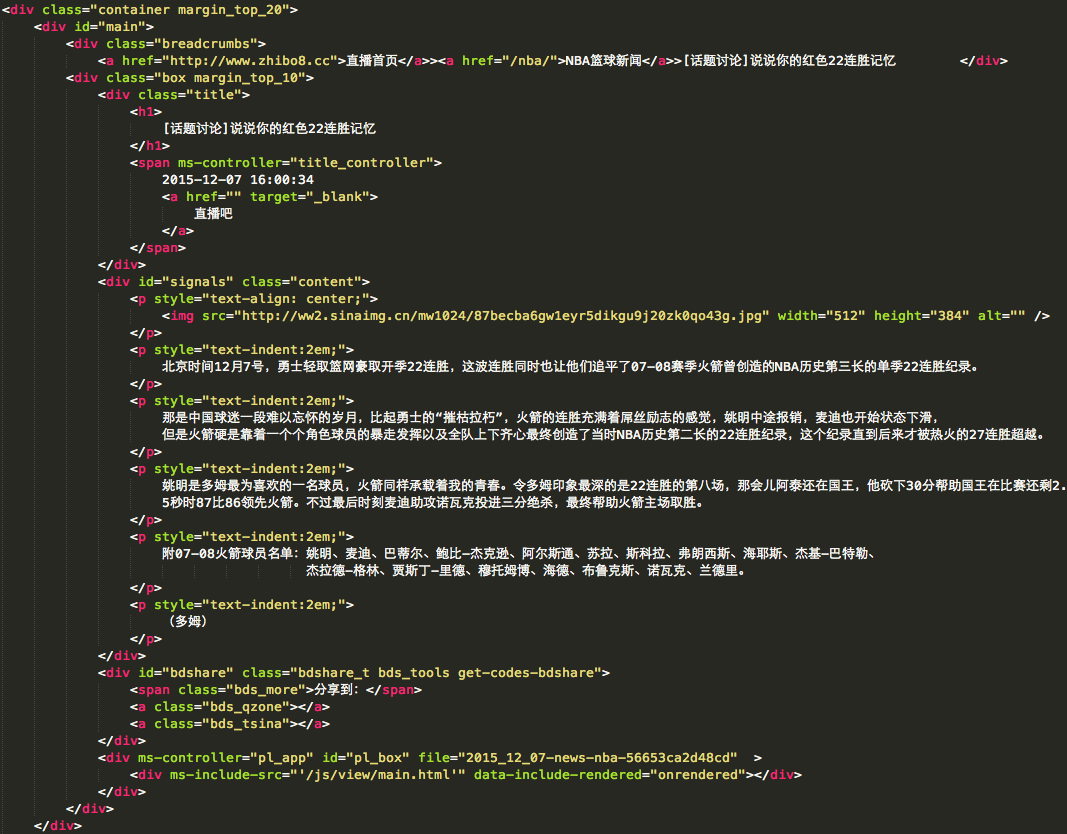
正如上文所诉，新闻网页的文本标签具有其特殊性，本小节将对其进行特征挖掘工程，确定后文方法中将要用到的文本标签特征，并详细讲解文本标签特征提取过程。

## 3.3.1文本标签特征工程

几乎所有的机器学习算法都依赖于特征选择工程，特征选择的优良将直接影响机器学习的结果，故本节的文本特征的选取将决定该文本提取方法的正确率。标签特征为HTML标签的开始标签、结束标签及其之间的标签内容的特征集合。通过对新闻网页结构解析，本文可以将新闻网页大致分为三大类：

1. 网页文本的组织结构。这部分标签主要包含<html>content</html>、 <head>content</head>、<body>content</body>等，它们主要是提供给浏览器进行网页解析使用，对网页正文文本提取几乎没有作用。
2. 网页中噪声资源。这部分标签主要包含广告信息、跳转链接信息、目录信息，如<script>、<noscript...>、<style...>、<img>等标签，对网页正文文本提取没有作用。
3. 网页正文格式，此部分信息包括文字概述、文章作者、文本信息、图片信息和少量的用户点评信息，这部分标签代表了网页作者对网页格式的特殊说明，将决定文本如何在浏览器中显示，这部分标签将是本分着重挖掘的部分。

进一步对各部分的文本特点进行分析发现，新闻文章信息部分中的文本信息标签具有中心化、连续性、层次性、长文性、左标签属性偏少和文本语言中中文字符数偏多等特点，而网页文本的组织结构、网页中噪声资源部分的标签偏向于零散、层次比较分散、相应的描述标签内容字符长度比较短。如下图简单的新闻正文HTML源码所示：



通过新闻文章正文标签的独有特点，本文主要挖掘以下六个标签特征：标签ID、标签内容字符串长度、标签内容中标点符号数、左标签长度、右标签长度和标签层次，其中标签ID依次递增，用于标签位置索引。

根据特征选择工程所选特征，本文在提取过程中定义了以下标签信息结构体，用于存储标签信息，将网页源码中有用信息存储在标签结构体列表中，方便后无的新闻识别及文本正文信息抽取。标签信息结构体如下：

struct lableFeature {

string lableName; //标签名

string lableContent; //标签文本内容

string lableLeftPartContent; //左标签内容

int lableBeginIndex; //左标签的开始位置

int lableEndIndex; //右标签的结束位置

int lableId; //标签ID

int lableContentLength; //标签文本长度

int lableLeftPartLength; //左标签长度

int lableRightPartLength; //右标签长度

int lablePunctNumber; //标签中标点符号的数量

//vector<string> lableAttributeVector; //标签存在的属性

int lableLevelNumber; //标签层次树

};

## 3.3.2文本标签特征提取方法

上小节的特征选择工程，概括出需要的提取的标签信息，结合后文新闻自识别方法中要用到标签坐标信息，本节将通过字符串处理的方式来获取预处理后的网页源码字符串中的所有标签信息。

标签具有层次性和嵌套性，例如下简单的HTML网页信息：

<div id=article>

Unknown information 1

<h1>

Title information

</h1>

Unknown information 2

<td id=article>

<a href=…..>

Link information

</a>

<td> ,。！</td>

<p>

Web page information 1

</p>

<p>

Web page information 2

</p>

</td>

</div>

在标签信息提取过程中如何确保正确的提取层次数和内容信息是一大难点，通过对下文的特定要求，在标签信息提取过程中针对特殊情况采用三种处理方式，分别为标签自补全处理、标签自修正处理和标签自过滤处理。

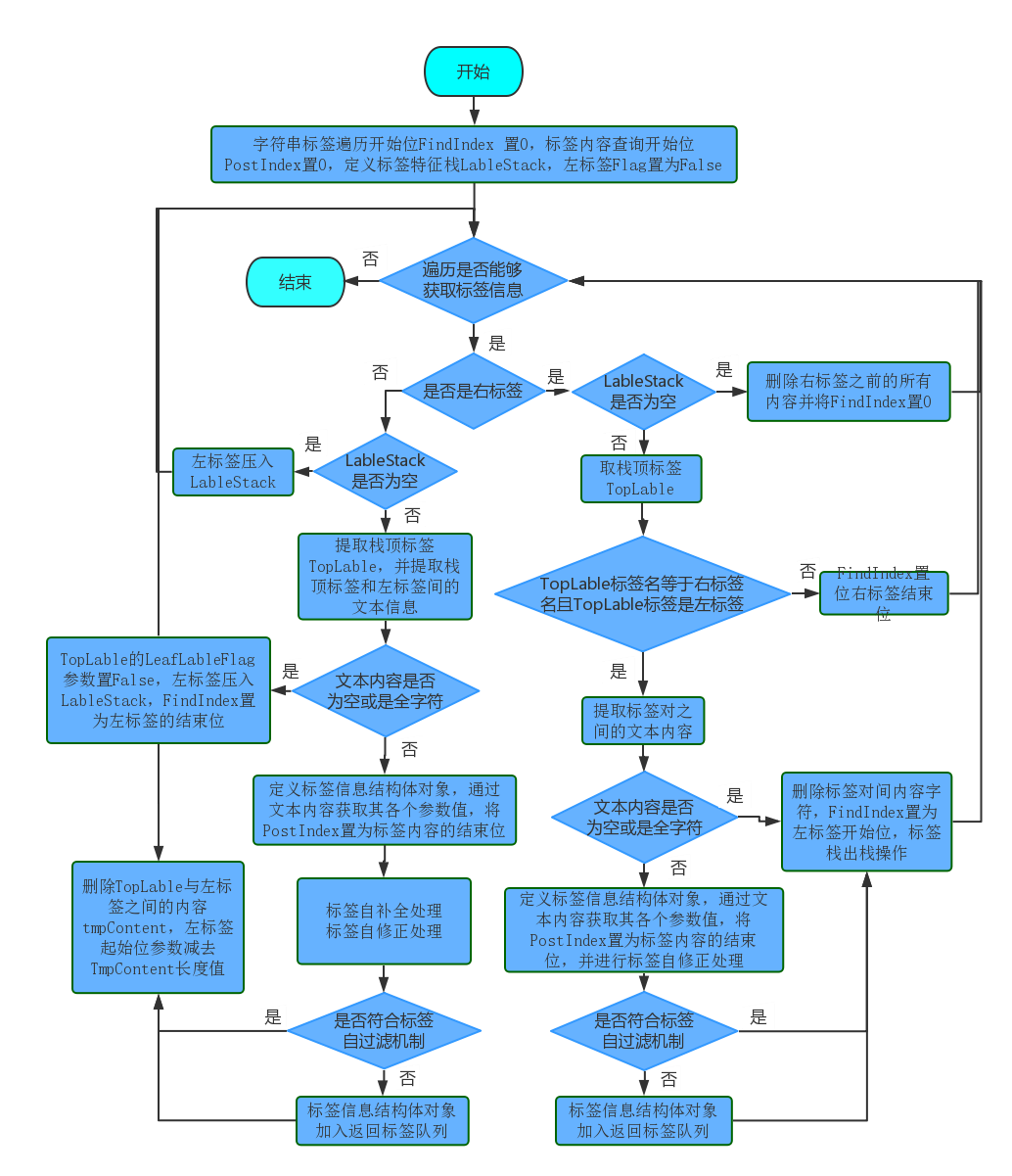
标签自补全处理是为了解决上图中“Unknown information 1”和“Unknown information 1”信息的提取及在全字符串中准确定位而添加的，当获取左标签div后，紧接着会获取左标签h1时，此时为了保证后续标签坐标的准确性，我们需要将“Unknown information 1”信息提取出来，做为一个div标签信息，就类似于在<h1>之前加入了”</div><div id=article>”，将左标签h1的起始坐标都剪去“Unknown information 1”的长度，如果“Unknown information 1”信息为空，则跳过这一步。

标签自修正处理是为了解决各个网站中网页自带的网站信息标签，如果标签内容中含有"All Rights Reserved"、"【免责声明】"、"all rights reserved"等任一字符串，方法中将自动将标签名改为a标签，因为在后续的机器学习过程中，a标签不参与训练，也就不会对正文簇选择造成影响，从而避免将网站信息规划到正文簇中。

标签自过滤处理是为解决空标签、标签内容全为标点符号或者且标签名为i且标签内容含有"http://….sohu.com…"的情况，此类标签信息对正文提取没有任何意义。在标签信息提取过程中将对此类标签进行过滤，并在预处理后的网页源码字符串中将其删除。

标签信息提取过程主要思想是借用数据栈，顺序从预处理后的网页字符串中提取HTML标签，若已经遍历完所有的标签，则退出。如果成功提取了标签，则分两种情况：一、提取的标签是左标签，判断栈是否为空，如果为空，栈将左标签压入标签栈；如果非空，则提取栈顶元素，判断栈顶标签和左标签之间的内容是否为空或是否是全标点字符，如果成立则不提取该部分信息，反之则提取该部分的内容为一个标签信息并加入标签信息队列中，并将该部分内容在原网页字符串中删去，再将左标签的起始坐标参数前移该部分内容的长度，最后讲左标签压入标签栈；二、提取的标签是右标签，如果栈为空，则跳过后续操作，继续遍历网页字符串获取下一个标签；如果栈为空，则取栈顶标签，分析栈顶标签和右标签之间的内容是否为空或者是否是全标点字符，如果是则在网页字符串中删除栈顶标签和右标签及其之间的内容，继续循环遍历获取下一个标签，将栈顶元素删去；如果不是则提取标签特征并加入标签信息容器中，在网页字符串中删除栈顶标签和右标签及其之间的内容，并对标签栈进行出栈操作，继续循环遍历获取下一个标签。

文本标签特征提取流程如下图所示：



标签特征提取方法流程图

## 3.4文本标签特征学习

本节主要介绍新闻正文提取方法中的特征学习方法。上节提取了文本标签特征，本节首先讲解标签特征向量的获取及各个特征在进行聚类过程中所占权重，然后通过不同种特征学习方法进行对比试验，确定适合本文的聚类算法，最后对算法进行总结。

## 3.4.1文本标签特征向量生成及权重选择

上文3.3节的文本特征提取过程，我们获取到网页文本的所有标签信息，通过这些有序列编号的标签如何获取和获取哪些标签信息将是本小节研究的内容。

每个HTML标签都有其对应的作用，有些是用于修饰网页显示文本的字体、有些是用于链接跳转、有些是用于列表展示、有些是为了提供文字输入等等，在新闻文本提取过程中我们最终是想获取新闻主题部分的标签，为了增加后文特征的学习速度，需要对标签进行筛选，去除那些对正文簇选择无用的标签，通过对比实验，本文根据实验结果将满足以下规则的标签添加到标签向量矩阵中，标签信息列表中数量最多的a标签信息对后文正文簇选择而言是冗余的，input的标签是缺点不可能出现在正文簇中的，li标签只对展示效果有影响，即使不参与聚类也可以很好的调整到正文簇中，ul标签则无有用信息，在特征聚类之前这几类标签将被过滤；新闻网页的标签长度普片相对于其他噪声标签偏长，在后文的聚类算法和文本簇选择过程中依据的是长文本性和多标点特点，再结合后文实践过程中对大量Bad case学习发现，标签文本长度小于10的标签会照常一定Bad case，所以本方法不会将标签文本长度小于10的标签用于文本特征聚类学习过程，另一方面去除这些标签也提高了聚类效率。

后文中将根据标签ID的连续性进行文本簇的经验性调整，标签特征向量中添加标签ID作为标识，本文将文本向量定为6维，分别为：标签ID、标签文本长度、左标签长度、右标签长度、标签文本标点符号个数、标签层次数。遍历标签信息列表，将满足上述规则的标签添加到标签向量矩阵中，获得一个N行6列的特征向量矩阵，然后将各维数据做归一化处理。

在对提取出来的标签特征向量矩阵进行大量的聚类实验后，发现聚类簇并不如我们直接在网页上通过视觉观察到的直观，各个聚类簇的特征不够明显，一些很明显的正文标签没有聚合到一个簇中，通过分析这种情况发生的原因，最终发现这类正文标签的层次数值特征和左右标签长度在聚类过程中占据了决定性作用，我们增加调整各个特征属性值的权重后进行聚类的对比实现，根据对比实验的结果确定将标签文本长度属性的权重增大，减小左右标签长度属性的权重。权重调整后再进行聚类，得到比较理想的聚类效果。

## 3.4.2 文本标签特征聚类方法

HTML网页的不同部分信息的特征不同，本节将通过特征聚类将所有标签进行聚类，然后针对各个聚类簇的特征做进一步挖掘。

聚类的目标是使同一簇文本标签的相似度尽可能的达到最大值，不同簇文本标签间的相似度尽可能的小。目前聚类方法有很多种，其各自又有不同的变形，根据基本思想的不同大致可以将聚类算法分为5大类：层次聚类算法、基于约束的聚类算法、机器学习中的聚类算法、分割聚类算法和用于高维度的聚类算法，其各自有各自的优缺点，为了避免算法的局限性和做一定的对比实验，本文选出了三种聚类算法来实现文本标签聚类，分别为： K-Means及其KMeans++聚类算法，二分K-均值聚类算法和AGENS层次聚类算法。

（1）K-Means聚类算法

算法大致流程：

1：随机选择K个点作为初始质心。

2：将每个点指派到最近的质心，形成K个簇。

3：重新计算每个簇的质心。

4：循环执行步骤2、3，直到聚类簇间的SSE值不变或者小于定值

5：退出算法。

K-Means聚类算法应用最为广泛，收敛速度快，适用于大规模的数据集，但是其结果好坏依赖于对初始聚类中心的选择、容易陷入局部最优解、对K值的选择没有准则可依循、对异常数据较为敏感、只能处理数值属性的数据、聚类结构可能不平衡。为了解决其初始聚类中心选择的随机性，方法中将K-Means升级为其的一个变种K-Means++算法，将随机选择K个初始质心改为通过各个特征值区间计算等距离取值的方式构造K个初始聚类中心。

介绍其缺点和聚类截图

（2）二分K-均值聚类算法

顾名思义，二分K-均值聚类算法就是每次对数据集或子数据集采取k=2的k-均值聚类划分，子数据集的选取则有一定的准则。二分K-均值聚类算法首先将所有点作为一个簇，第一步是然后将该簇一分为二，之后的迭代是：在所有簇中根据SSE选择一个簇继续进行二分K-均值划分，直到得到用户指定的簇数目为止。根据SSE选取继续划分簇的准则有如下两种：一、选择哪一个簇进行划分取决于对”其划分是否可以最大程度降低SSE的值。这需要将每个簇都进行二分划分，然后计算该簇二分后的簇SSE之和并计算其与二分前簇SSE之差，此时SSE必须下降，最后选取差值最大的那个簇进行二分。二、在所有簇中选择SSE最大的簇进行划分，直到簇数目达到用户指定的数目为止，算法过程与准则一类似，区别仅在于每次选取簇中SSE最大的簇。

本文选择第二种准则进行簇的选择，算法流程如下：

1：将所有数据看成一个簇。

2：运用准则2选取待划分的聚类簇。

3：将第2步选出的簇运用KMeans进行二分，用获得的两个子分类簇替换被划分的分类簇。

4：循环执行步骤2、3，直到簇数目等于K。

5：退出算法。

介绍其缺点和聚类截图

（3）AGENS层次聚类算法

AGENS算法是自底向上的层次聚类，首先计算任意两个向量的距离，然后对所有距离进行排序，接着系统聚合具有最小点对距离的两个簇，直到达到用户输入的聚类数目，该方法拥有较高的聚类精度但算法复杂度较高，适合于小数据量的聚类情况。

算法流程如下：

1：计算所有点对之间的距离。

2：选取不同簇中具有最小距离的两个点，聚合两个点所在的簇。

3：重复步骤2，直到达到聚类数目K。

介绍其缺点和聚类截图

综上三种聚类算法，其各自有各自的优缺点，在经过大量数据对比实验后，选择AGENS层次聚类算法对后文的文本正文提取正确率提升最大也最稳定，虽然AGENS层次聚类算法复杂度偏大，但是由于网页文本标签特征向量数普片不超过1000个，所以在最求高正确率的情况下，AGENS层次聚类消耗的时间在可接受范围之内。

添加结果对比表格

## 3.5 网页正文簇选择及标签经验性调整

以上三个小节已完成网页标签预处理及修复、标签特征提取及清理和网页特征聚类，本节将根据3.4节的聚类结果选择出网页正文簇，确定网页正文标签名和标签层次，再根据新闻网页正文标签特点进行正文簇内标签的经验性调整。

## 3.5.1网页正文簇选择策略

网页正文簇定义为在网页文本标签特征进行层次聚类结果中和网页正文特征最相似的分类簇，此分类簇中的标签可以很好的描述网页正文标签特征。

上小节介绍了标签特征向量的聚类，最终选择了AGENS层次聚类算法做为特征聚类方法。聚类算法实现部分中包含聚类簇中心向量生成接口，该接口是基于加权平均的思想去获取中心向量，中心向量可以很好的描述各个分类簇中标签的特征。本方法中正文簇的选择主要是通过挖掘各个聚类簇中标签特征，选取最符合正文标签特征的簇做为正文簇，方法中有两种选择策略，如下所述：

其一是根据各个聚类簇中心向量的标签文本平均长度与平均标点数之和来确定，取和最大的聚类簇做为正文簇，在网页文本标签特征聚类完成后，各个聚类簇都有一个聚类中心向量，在标签层次聚类过程中K值的选取会影响到最终的聚类结果，当新闻文本标签数很少且特征不是特别明显的时候，K值选取就变的尤其重要，当K值很大时才能很好的将正文标签聚到一个簇中，当K值选取不合理时，则包含文本标签的簇中会含有很多噪声标签，此时根据聚类中心向量进行正文簇选择，其选取正文簇的错误概率将增大。

其二是根据新闻网页正文标签文本的长文性和具有分段分句的特点，各个聚类簇中文本长度和标点符号数之和作为判断依据，取各个聚类簇中文本长度和标点符号数之和最大的五个标签信息，再计算取出的Top标签的加权平均值，获取对应的加权向量，如果某个聚类簇中的标签特征向量数量小于5个，则取该聚类簇的中心向量做为加权向量，最后取加权向量中文本平均长度与平均标点数之和最大的聚类簇做为正文簇。

增加两种策略选择正文簇的对比实验，实验数据集为500个新闻网页，数据集中包含长正文的网页、短正文的网页、正文标签数多的网页和正文标签数少的网页，这几种网页的比较区域等比例；实验结果如下：

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 正文簇选择策略 | 新闻网页数量(个) | 文本簇选择正确数(个) | 文本簇选择错误数(个) | 正确率 |
| 策略一 | 500 | 483 | 17 | 96.6% |
| 策略二 | 500 | 498 | 2 | 99.6% |

通过实验结果，我们发现策略二相对于策略一的效果要好，在方法中也将选择策略二进行正文簇选择。通过对正文簇选择错误网页的标签分类簇分析发现，策略一对短正文网页和正文标签数偏少的网页正文簇选择错误率要高很多，策略二中出现的两个错误网页为多图网页或一句话正文网页，在方法的后续步骤中会对这类特殊网页再进行一次过滤。

添加错误截图

## 3.5.2 网页正文簇标签经验性调整

为什么要进行正文簇标签调整，原因大致有以下三点：一、一些非正文文本标签特征和正文标签特征比较接近，在聚类过程中也会将这类标签聚合到正文簇中；二、网页正文标签的特征也不完全相似，可能有些正文标签的文本长度很短、标点符号数值也很下，在聚类时候可能会分聚合到非文本簇中；三、在特征聚类之前，我们将对文本簇选择影响不大的标签过滤了，但这些标签有可能是嵌套在正文标签内部的标签，用来改变文本的格式或添加超链接，例如a标签、li标签和文本长度小于9的标签等，而在网页正文中会出现对关键词添加超链接的情况，既存在a标签，网页正文中也会出现数据列表，即存在li标签，也会出现长度小于9的标签。故上节获取的网页正文簇中的标签只混杂的，有些是正文标签、有些不是正文标签，所以需要对正文簇进行调整。

正文簇标签调整的主体思想：在已知网页正文簇中存在大量的正文标签的情况下，可通过这些的正文标签去确定网页正文标签名，标签层次和标签ID分布位置，再根据新闻网页正文标签的连续性和中心化特点，从已知标签ID进行中心扩散。

正文簇标签调整可分为以下几个步骤：

1、获取正文标签名、标签层次数和若干个确定为正文标签的标签ID，把这些标签ID作为中心扩展标签ID组。

2、根据正文标签名、标签层次数和中心扩展标签ID分别进行正文标签中心扩展，选定一个正文ID为中心向前后扩展正文标签范围，形成一个正文标签子中心块。

3、合并正文标签子中心块及非正文块过滤

4、合并并过滤后的所有子中心块包含的标签ID定为新的正文簇。

第一步获取正文标签名、标签层次数和和若干个确定为正文标签的标签ID，根据正文簇中正文标签的长文本特征及正文标签的混杂性，取正文簇中标签文本长度最长的5个作为判断正文标签的依据标签，在提取过程中要排除一部分不可能为正文标签名的标签，例如标题标签h1、h2、… 、h9标签、td标签和a标签等，如果正文簇中标签向量个数小于5个，则取全部标签向量为依据标签，从依据标签中取标签名出线次数最多的标签名作为正文标签名，如果存在次数相同情况，则取文本长度最长的标签名作为正文标签名。此过程中对依据标签数目只有两个且两个标签名不同的情况要做特殊化处理，若第二个依据标签ID接近总标签ID数，则定第一个依据标签的标签名为正文标签名；否则根据两个依据标签的标签ID与其前后标签的连续性进行判断，两个标签的连续性定义为前一个标签的文本结束坐标紧接后一个标签的文本开始位，连续数多的标签名定为正文标签名。

确定正文标签名后，将提取的依据标签中标签名等于正文标签名的标签ID定为确定的正文标签ID。如果正文标签名对应的标签层次数都相同，则定这个数值为正文标签层次数，如果各不相同，则将正文标签层次数置零，因为有少部分网页的正文标签不具有同层次性，其正文标签层次数呈现递增的趋势，将正文标签层次数置零是为了在后续正文标签调整过程中不用同层次性这个特征。

第二步根据正文标签名、标签层次数和已知正文标签ID分别进行正文标签中心扩展。此部分内容依据是大量观察新闻网页正文的结构特征，根据其特征归纳出正文簇中心扩展方法，该方法的中心思想是将已知正文标签ID做为中心，把正文标签的层次性、连续性特征和网站描述内容的特定性作为中心扩展的约束条件，为了尽可能多的提取正文标签，在扩展中针对不确定情况和特定标签加入了松弛机制，例如允许一定数量的不确定是不是正文标签的标签加入到正文标签子中心块中，而不是直接结束中心扩展，扩展后的标签ID集合形成一个正文标签子中心块。由于中心扩展时加入了松弛机制，需要对子中心块的首尾进行一定的过滤，将首尾标签名不等于正文标签名、不连续的标签ID删去，最后返回正文标签子中心块。

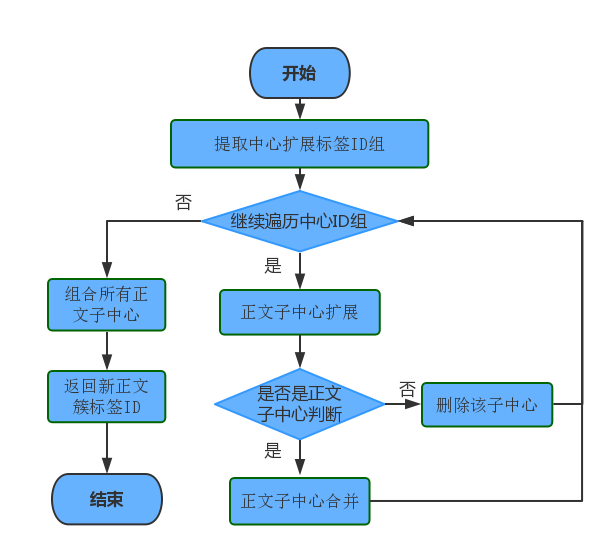
第三步合并正文标签子中心块及非正文块过滤。通过第二步的正文中心扩展，理论上讲第一步确定了多少个正文标签ID就会形成多少个正文标签子中心块，但有些子中心块的ID是连续的，实现过程中将连续的子中心块合并为一个。由于第一步判定为正文标签ID只是通过特征聚类后按照一定规则获取的，会出现误判情况，这种情况下通过误判的“正文标签ID”进行中心扩展获取的子中心块将不是真正的正文标签子中心块，故在正文标签子中心块合并的过程中需要添加适当的过滤策略。

非正文标签子中心块的判定及过滤策略：

1. 针对合并后只有一个正文标签子中心块的情况。若子中心块内标签ID数量小于4个且标签ID分布在全标签信息的尾部，此时求取子中心块中标签文本的平均长度，如果标签文本平均长度很短且网页的总标签数很大时，判定这子中心块为非正文标签子中心块，并将该子中心块清空。
2. 针对合并后存在多个正文标签子中心块的情况。这类情况下，首先判定并清理单标签ID的子中心块，遍历所有子中心块，若某子中心块只有一个标签ID，如果这个标签ID与其前后的标签ID不具有连续性，则将此子中心块判定为非正文标签子中心块，并将此子中心块清除。其实判断最后一个子中心块，判断策略是根据最后一个子中心块标签ID在全标签信息的位置和其与前一个子中心块相对位置而定。

第四步将过滤后的所有子中心块包含的标签ID定为新的正文簇，此时的正文簇中的所有标签ID是待提取正文内容的标签，其还要经过新闻网页自识别方法进行新闻网页判定。

网页正文簇标签经验性调整的流程图如下图：



## 3.6非新闻网页自识别方法研究

上文3.5小节对正文簇进行经验性调整后，能够获得正文标签ID的集合，本节将通过挖掘正文簇内标签的特征及相互关系来进行非新闻网页自识别。

非新闻网页包含网站引导页面、大部分BBS论坛页面、部分多图且无正文的网页页面或者多图有少部分对图片描述文字的页面和多数问答类型的页面等等，这类页面有共同的特征，如网页文本信息几乎没有或很少、问答类页面的题目描述很短而回答很多、BBS论坛帖子的主题很短而回帖很多，在新闻采集过程中要对这类网页做相应的过滤，才能保障提取到的网页内容尽可能多的是新闻网页内容，保障了采集信息的可信度。

网页信息获取是通过对网页URL模仿浏览器请求获取网页信息。URL是互联网统一资源定位符，完整的URL由五个部分组成：协议、服务器地址、端口号、路径和查询参数，例如美国篮球巨星科比的维基百科页面的URL：https://zh.wikipedia.org/w/index.php?title=%E7%A7%91%E6%AF%94%C2%B7%E5%B8%83%E8%8E%B1%E6%81%A9%E7%89%B9&printable=yes。新闻采集依赖于蜘蛛网络爬虫，不断的从网站获取网页，解析其网页正文，提取网页中的超链接信息加入到待爬去网页的URL列表中，再根据设定的爬虫策略选择下一个该爬取的URL。每个网页中都会含有众多的超链接URL，通过URL组成部分可知，我们很难分辨出哪些网页是新闻网页那些网页不是新闻网页，所以在将爬虫过程中解析到的url加入到待爬去网页URL队列时，无法保证待处理的URL队列中全是新闻网页的URL。通过对URL地址特点的研究，对于一部分非新闻网页的判断，可以通过URL特征进行预判断。静态网页的URL一般以.htm、.html、.shtml和.xml等形式为后缀结尾，大部分非新闻网页都不是静态页面，通过判断URL是不是静态网页可以对非新闻网页进行一层过滤，但是对于一句话新闻网页或多图无正文网页来说，将没有合适的办法进行过滤，本节将在网页提取过程中特征工程的基础上，根据提取出的网页文本标签的特点及布局进行非新闻网页的自识别，对不符合新闻采集系统要求的网页进行适当过滤。

## 3.6.1 非新闻网页自识别方法介绍

通过上文可知，新闻网页的正文结构具有中心性、连续性和长文本性质，进一步挖掘网页正文簇内正文标签名对应的标签数、所有正文簇内标签文本平均长度、正文标签名标签文本平均长度、格式标签数、链接标签、标题标签数、div标签数、正文中心数和标签聚类等数据分布情况，通过大量的实验结果获取这些数据的大致分布，设定相应的判定阀值进行非正文判断。根据新闻采集系统应用的特点，比如舆情分析、数据监控等，大千网站中，对于多图新闻网页、短正文的新闻网页等特殊网页，我们本着尽可能多的采集新闻原则，即将完全可确定为非新闻网页的网页过滤，对于一些非新闻网页但它们具有新闻网页的绝大部分特点，本文将把这类网页划分为新闻网页；对于一些不完全具备新闻网页特点但其实际上是新闻网页，我们将并其定为新闻网页。

非新闻网页自识别方法主要思想是基于前文获取的正文簇内标签信息特点挖掘，本着尽可能少误判的原则设定相应判定阀值，进行非新闻网页的判别。该方法步骤大致分为以下三步：

第一步正文簇内标签信息挖掘。通过加权计算获取正文簇内标签文本的平均长度、正文标签名对应的标签文本平局长度、正文标签名对应的标签数、修饰文本格式的标签数、超链接a标签数、标题h系列标签数、容器标签div标签数、其他标签数、正文簇标签的子中心数、正文簇标签ID非连续的标签数、相邻标签ID的间隔距离大于40的个数和标签文本长度最长的标签ID。

第二步设定各特征参数的判定阀值及调整。该方法本着尽可能少误判的原则，在实现初期将各个阀值设定的都很放松，增加验证性实验计算误判率，针对性的分析误判网页正文簇标签特征值，对相关阀值进行调整，不断的进行对比实验，循循渐进的修改，直到取到所能接受误判率范围的阀值。

第三步根据阀值和判定策略进行非新闻网页判断。非新闻网页的判定策略是根据新闻网页结构特征的中心性、长文本性和连续性特点而设定。未经判定的网页默认为是新闻网页，将步骤1提取出的各参数按照如下顺序进行判别：

1. 若正文簇中标签ID为空，则判定为非新闻网页，直接跳过其他判定策略。
2. 正文标签名标签文本的平均长度小于对应阀值，且相连标签距离大于40的标签个数大于对应阀值，则判定为非新闻网页。
3. 标题h系列标签数大于对应阀值，且总标签文本的平均长度小于对应阀值，则判定为非新闻网页。
4. 超链接a标签数大于正文簇中总标签数的一半，则判定为非新闻网页。
5. 相邻标签之间距离大约40的个数大于正文簇中总标签数的一半，则判定为非新闻网页。
6. 正文标签名标签文本的平均长度小于60，且正文标签名对应标签数小于等于4，则判定为非新闻网页。
7. 如此时网页已被判定为非新闻网页，若文本簇标签中最长的标签文本大与对应阀值，则判定为新闻网页。
8. 标题h系列标签的个数大于正文簇中总标签数的一半，则判定为非新闻网页。
9. 正文簇标签ID非连续的标签数大于12或大约正文簇中总标签数的2/3，则判定为非新闻网页。

10、正文簇标签的子中心数大于3个或者等于正文簇中总标签数，则判定为非新闻网页。

11、若正文簇中只有一个标签，且标签文本长度大于设定阀值，则判定为新闻网页。

经过非新闻网页自识别方法判定后，可以确定该网页是否是非新闻网页，为后文是否进行文本簇内标签的文本信息提取并存储提供判断依据。

## 3.6.2 非新闻网页自识别方法效果评估

在进行本实验前要进行数据采集，为了保证实验结果的可信度，我们首先选择20个新闻门户网站的主页URL，其次把各个主页URL做为python爬虫脚本的初始种子，不间断的爬取该站内的500个网页，然后人为的进行非新闻网页的识别，将新闻网页URL和非新闻网页URL分别写入对应的文件，这样数据集中就存储了从20个网站中提取的10000个网页URL，并进行了非新闻网页和新闻网页的区分作为样本集合。

实验中所使用的数据集主要包含两个部分，其一从获取到的新闻网页中随机抽取1500个新闻网页URL，其二从获取到的非新闻网页中随机抽取500个非新闻网页URL。

通过实验结果统计分析，主要统计如下表所示：

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 实际类型\判定类型 | 总数量(个) | 新闻网页(个) | 非新闻网页(个) | 正确率 |
| 新闻网页 | 1500 | 1483 | 17 | 98.87％ |
| 非新闻网页 | 500 | 42 | 458 | 91.60% |

对新闻网页判定的统计发现，在1500个新闻网页中有17个被误判为非新闻网页，细致观察这17个误判的新闻网页后，发现这类网页具有以下特点：一、正文标签名对应的标签数不大于两个且标签文本长度很短，二、网页的正文簇中标签的文本偏短，且存在大量对文本内容注释的超链接a标签和制表格标签。

再进一步统计可得，新闻网页误判网页中一部分网页属于一句话短正文新闻网页，从新闻采集的目的性来看，这类网页即使提取出来也无用处，故不把其归化为误判；一部分属于多图网页，其网页的正文就是一句对图片描述的短文字，这类网页可认为非新闻网页，故不把其归化为误判；剩余的网页确实为新闻网页，其正文簇标签中包含大量a标签和制表格标签，本方法无法正确的判定，但这类新闻网页的数量很少，故新闻网页判定的实际误判率小于0.5%，在新闻采集过程中可被接受。

对非新闻网页判定的统计发现，在500个非新闻网页中有42个被误判为新闻网页，细致观察这42个误判的非新闻网页后，也发现这类网页有以下特征：一、网页正文簇中存在标签文本较长的标签；二、正文簇标签中有多个子中心块；

再进一步统计可得：非新闻网页误判的网页中有部分是网站引导页面，其中页面中的跳转链接标签对应的文章标题很长，且跳转链接标签具有连续性和中心化特点，在标签特征挖掘时将这部分标签选入正文标签簇，故判错；其中有部分是bbs论坛网页，其论坛主题很短，但是其跟贴的标签文本长度很长，特征挖掘过程中将该网页中跟贴标签划分到正文簇中，故判错；第三部分网页无规律可言，故判错，这部分大约占到3%左右。对前两部分误判网页来说，本着尽可能多的提取网页正文的原则，这两类误判网页被识别为正文也是可行的，毕竟这部分内容可以很好的描述该网页的内容。

总结上述实验，通过对误判网页进行分析，最终只有0.5%的新闻网页会被误判为非新闻网页，大约3%的非新闻网页被判断为新闻网页，97％的非新闻网页可正确判定，效果符合新闻采集的尽可能多的采集网页信息的需求。

误判网页源码截图

## 3.7网页正文簇标签内容提取

本节是网页正文提取方法的最后一步，由于正文文本内容及格式的多样性，本文的提取方法只针对提取文本文字信息，其表格显示形式和图片将不做提取。在提取文本簇文本之前，需要先对该网页进行非新闻网页的判定。在3.6节获取网页正文簇后，进行非新闻网页自识别判断，如果判定为非新闻网页，则不进行标签文本提取及存储，直接反馈改网页是非新闻网页；若判断为是新闻网页则进行标签文本提取和存储。

本章的3.3小节在提取文本标签特征提取时，从文本标签特征结构体可发现，提取特征的同时已将标签内容提出，在文本标签特征提取过程中标签ID是连续递增的，标签文本内容的位置在网页字符串中也是相对有序，在提取网页正文时，只要按照正文簇中标签ID的顺序提取出对应的标签文本内容，再依次组合在一起，即可得到网页正文内容。

## 3.8 实验及文本提取性能评估

在该部分，主要介绍了实验过程中所用的数据集，实验结果的评价标准和性能分析。给出了我们基于文本标签特征挖掘的网页正文提取方法在这些数据集上的实验结果，并于前人在网页文本提取方面的提取结果进行比较。

本节将进行两个实验，实验一：在文本提取方法中不添加非新闻网页识别过程，对所有测试网页都进行正文提取并判断提取是否正确。实验二：在文本提取方法中增加非新闻网页识别过程，若判断该网页为非新闻网页，在进行网页正文提取错误分析过程中，将判断为非新闻网页且网页实际确实可认为是非新闻网页规划为提取正文正确的范畴内。

以下实验的硬件配置如下图：



## 3.8.1实验数据集与提取性能评估指标

实验分析中所采用的评估指标：正确率和提取效率。

1. 正确率是样本集中正确提取出正文的网页数所占样本集的比例。正确提取网页正文的定义：新闻网页正文内容的95%被抽取出来，且基于方法提取出来的正文中实际网页正文的内容篇幅不小于85%。即假如网页中正文长度为1000个字符，这本方法至少要提取出来950个字符的正文，且假如基于方法提取出来的网页内容长度为1000个字符，此时至少有850个字符是实际网页的正文部分。
2. 提取效率是处理一个网页需要多长时间，用单位时间内处理网页的数量来衡量。

实验一中使用的测试数据主要分为两个类：

（1）中文网页测试集：该数据集包括30个不同的中文新闻网站的网页信息，每个网站各包含100个网页，人为的调整该网站网页中长正文文本和短正文文本的网页比例，调整后比例为9:1，这也是为了避免测试数据中长文本正文网页过多而影响实验结果的可信度，并在短正文文本中调整一些多图新闻网页，从而构成总数位3000个网页的中文网页测试集合。。

（2）英文网页测试集：该数据集中包含10个不同的英文新闻网站的网页信息，每个网站各包含100个网页，调整网页集中长正文文本和短正文文本的网页比例为9:1，构成总数位1000个网页的英文网页测试集合。

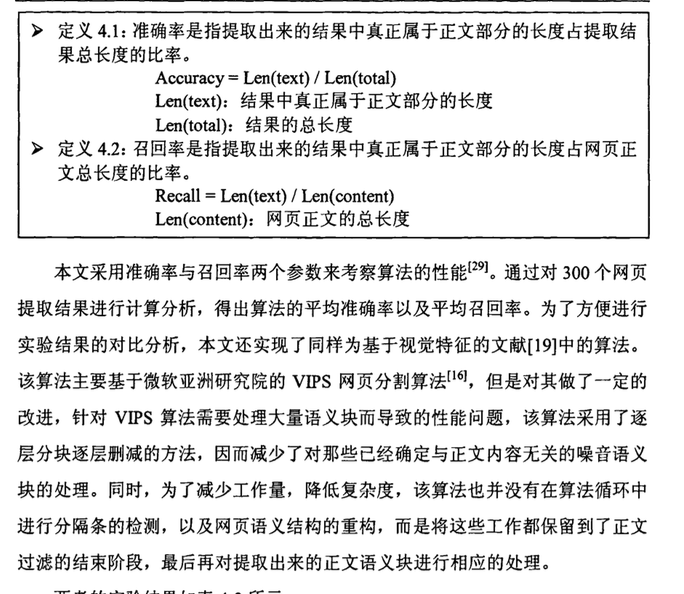
实验二中使用的测试数据主要分为两类：

（1）中文网页测试集：该数据集包括30个不同的中文新闻网站的网页信息，每个网站各包含100个网页，其中90个新闻网页，10个非新闻网页，新闻网页集又人为的调整长正文文本和短正文文本的网页比例，调整后比例为8:1，并在短正文文本中调整一些多图新闻网页，这也是为了避免测试数据中长文本正文网页过多而影响实验结果的可信度，从而构成总数位3000个网页的中文网页测试集合。

（2）英文网页测试集：该数据集中包含10个不同的英文新闻网站的网页信息，每个网站各包含100个网页，其中90个新闻网页，10个非新闻网页，构成总数位1000个网页的英文网页测试集合。测试集内的网页调整同中文网页测试集的调整。

## 3.8.2 实验结果及评估分析

（重写：参考基于视觉语义块的网页正文提取算法研究——浙大研究生毕设）



实验一的实验结果如下表： 表中加上处理时间、及效率值

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 数据集 | 网页数（个） | 正确结果（个） | 错误结果（个） | 处理时间（s） | 处理效率（个/s） | 正确率 |
| 中文网页 | 3000 | 2933 | 67 | 126 | 23.8 | 97.77% |
| 英文网页 | 1000 | 972 | 28 | 41 | 24.3 | 97.2% |
| 中英网页 | 4000 | 3905 | 95 | 167 | 24.0 | 97.63% |

通过图表我们可以发现，中文网页测试集的正确率要略为比英文网页测试集的正确率高，可能是因为在特征挖掘和策略调整时用到的参数值设定大多数都是通过中文网页的测试结果设定，中文网页和英文网页的编码格式也不同；英文网页的处理效率要比中文网页的高，可能是由于网页的大小同造成一定的效率偏差，也有可能是因为英文网页的广告、相关阅读的噪声标签相对较少，在标签特征学习的时候相对要速度快。

再对出错的网页进行针对分析,我们发现,大约一半出错的网页有一个共同特征,它们正文标签包含的信息都很短,或者只有简单的一句话,或者是包含少量文字说明的图片新闻，这样的网页中,因为其所含的信息量较少,所有通常的新闻网页的特征完全不同，如果排除这类网页的错误，总体的正确率可以达到99%。在效率方面，单线程每秒能够处理24个网页，效率也是很可观的。

实验二分析：考虑到实际运用中，任意给一个网页，通过URL无法判定是否是新闻网页，如果网页是非新闻网页，其无正文，故也就不可能正确提取正文，所实验二中加入了非新闻网页的判断，在提取过程中添加非新闻网页的识别过程，如果是非新闻网页则不提取正文内容，反之则提取。在正文提取正确性判定上本文将对特殊网页做以下调整：

特殊情况一：一些新闻网页正文标签信息数量很少且信息偏短、或者正文只有简单的一句话、或包含少量文字说明的图片新闻，这类新闻网页将被判定为非新闻网页而不提取其正文内容。鉴于这类网页的特殊性且网页正文信息量少，我们将其视为非新闻网页，在判断提取是否正确时，将其判定为提取正确的范畴。

特殊情况二：一些非新闻网页其网页内部分标签文本偏长且具有中心性，而且这部分标签的文本能够概括该网页的主题大意，在非新闻网页判定过程中会将其判定是新闻网页，针对这种特殊的非新闻网页，我们在判断正文提取过程中是否正确时，将其判定为提取正确。

特殊情况三：有非新闻网页不需要提取正文，所以将非新闻网页且被正确判断为是非新闻网页的这部分网页判定为正文提取正确。

实验二的实验结果如下表：表中加上处理时间、及效率值

备注：

Num1:新闻网页中正文被正确提取数量 Num2:满足特殊情况一的数量

Num3:满足特殊情况二的数量 Num4:满足特殊情况三的数量

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 测试数据 | 网页数(个) | Num1(个) | Num2(个) | Num3(个) | Num4(个) | 时间（s） | 效率值（个/s） | 正确率 |
| 中文网页 | 3000 | 2667 | 18 | 16 | 275 | 130 | 23.1 | 99.2% |
| 英文网页 | 1000 | 883 | 9 | 5 | 92 | 42 | 23.8 | 98.9% |
| 中英网页 | 4000 | 3550 | 27 | 21 | 367 | 367 | 23.3 | 99.13% |

通过实验二的结果可发现，在增加非新闻网页自识别过程后，其正确率趋近与我们的预期值，也更加符合新闻采集应用。其中中英文网页的正确率相比于不加非新闻网页自识别过程而言要提升很多，非新闻网页的提取错误率得以修正。效率方面相对要低一点，因为非新闻网页自识别方法的判别过程需要一定的处理事件，但效率降低有限。

从上面的实验结果可以看出,该基于文本标签特征挖掘的网页正文提取方法具有通用性的同时,保持了较高的正确率。在增加非新闻网页自识别过程后，不仅能够高可靠性的识别非新闻网页，还避免了非新闻网页正文提取错误的发生，使得方法在测试数据集上的表现更加优秀。在实际教研室工程项目运用中，我们对来自100多个网站的网页进行正文提取，其正确率依旧保持在98%以上，处理效率也有一定的保障，这说明该提取方法具有实用性和通用性。

## 3.9本章小节

本章探索研究了Web新闻网页的正文特点，通过挖掘正文文本标签的特点，提出了基于文本标签特征挖掘的网页正文提取方法，章节中详细介绍了本方法的主要思想、特征挖掘工程、特征学习方法、正文簇经验性调整等具体过程，在相应小节中加入了大量实验，并对本文方法进行实验分析，验证其可行性和超高的正确率。在基于文本标签特征挖掘的网页正文提取方法的基础上，设计了一种非新闻网页自识别方法，并进行大量实验分析，验证方法的优越性。

# 应用实例－智能Web新闻采集系统设计

上一章讲述了新闻文本提取方法和非新闻网页自识别方法，本节将介绍智能Web新闻采集系统中所涉及的网络爬虫、网页解析和新闻网页识别模块。

## 4.1 概述

新闻采集系统主要功能是根据用户自定义的任务配置，批量而精确地抽取目标网络媒体栏目中的新闻或文章，转化为为结构化的记录(标题，作者，内容，采集时间，来源和分类等)，保存在本地数据库中，用于内部使用或外网发布，快速实现外部信息的获取。现有的采集系统对特定的网站设置相应配置文件，不具有通用性，智能Web新闻采集系统的智能提现在系统的通用性，其可以采集任意新闻网站的网页，并对新闻网页进行智能化解析，不用对任何网站进行特殊配置，做到真正的无模版化。

智能新闻采集系统的基础是网络爬虫，在爬虫获取网页信息后进行网页解析，识别其是否是新闻网页，获取网页的标题、正文和相关链接等信息，再根据爬虫策略将相关链接插入爬虫URL列表，依次不断的在全网爬取各大网站新闻，直到网络爬虫的停止条件为止。

## 4.2网络爬虫设计

网络爬虫是整个新闻采集系统的基础，它根据 HTTP 协议向 Web 网站服务器 发送请求，Web 网站服务器接收请求并对请求进行分析,最后向网络爬虫返回 URL 的报文,从而实现网络爬虫对网页的抓取。运行网络爬虫时,只要给出 URL 种子网址,网络爬虫就可以按照指定 的抓取规则沿着网页上的超链接进行漫游,收集网页上的信息,直到达到了网络 爬虫的停止条件为止。一个设计良好的网络爬虫应该具有以下的技术: 初始种子的选择、爬虫的搜索策略、避免网页重复收集、网页的预处理、网页的识别策略、中文分词策略、多线程抓取网页。接下来将详细讲解爬虫的初始种子选择、搜索策略、中英文分词策略。

## 4.2.1初始种子的选择

网络爬虫是通过构造网络请求去获取特定网页的数据，网络爬虫的初始种子就是网络爬虫的抓取入口。众所周知,互联网上的网页 都是以超链接的形式进行关联的,每个 URL 都是互联网的一个节点,利用六度空 间理论,我们可以知道,从某一个节点出发,经过若干次的跳转,是可以和其他的任意的节点取得联系[33]。基本上,互联网中所有的节点都是互联的。

由于互联网规模非常庞大，在想获取特定相关信息的前提下，初始种子的选择将至关重要，如想获取整个互联网中新闻网页信息，可以通过各新闻门户网站为入口，

根据新闻采集系统的作用而定初始种子选择策略

## 4.2.2爬虫的搜索策略;

互 联 网 可 以 看 成 是 一 个 巨 大 的 “ 图 ”, 每 个 U R L 都 可 以 看 成 是 图 上 的 “ 节 点 ”, 页面中的超链接则是图上的“有向边”,爬虫的搜索策略其实就是对互联网这个图 的遍历策略。爬虫的搜索策略是指一种事先已经设定好的、比较合理有效的爬行 搜索方式,网络爬虫对网络上的节点进行有序地抓取。爬虫的搜索策略较为通用 的有深度优先搜索和广度优先搜索、最佳优先搜索这三种方式。

深度优先搜索策略类似于图的深度优先遍历,就是让网络爬虫在抓取网页的时候尽量往网页的更深层次挖掘,网络爬虫尽量往远处的 URL 搜索,直到无法搜 索了再按照原路返回。网络爬虫在使用深度优先搜索策略进行网页抓取中,应当 对网页进行去重。

宽度优先搜索策略类似于图的宽度优先遍历,也称为层次优先搜索。宽度优 先搜索策略需要已访问队列(又称 Visited 队列)和未访问队列(又称工作队列,Todo 队列)的支持,其更适合于分布式网络爬虫,启动多个网络爬虫,每个爬虫负责一 层。

最佳搜索策略是按照一定的网页分析算法,判断网页的主题相关性,选出 Todo 队列中主题相关性比较高的 URL 优先进行抓取。Fish-Search(鱼群算法)和 Shark-Search(鲨鱼算法)是主要的最佳搜索策略算法。鱼群算法是让网络爬虫抓取 网页的过程看成是鱼群觅食的过程,鱼代表网站 URL,食物代表主题相关网页[34]; 鲨鱼算法是在鱼群算法基础上加入 URL 与主题的相关度,是对鱼群算法的改进[35]。 本文使用的网络爬虫搜索策略就是最佳搜索策略。

本新闻采集系统将采用聚焦网络爬虫，

分词＋语义相似

## 4.2.3中文分词策略

NLPIR汉语分词系统

## 4.2.4网页的预处理（URL）

一般地,网页的预处理包括对抓取下来的网页进行链接提取、文字抽取、分 词等。网络爬虫抓取的文件格式不一,包括 Word、PDF、Excel 等,Java 中的相 应的 Jar 包提供了对这些文件相应的处理接口。网络爬虫需要根据文件的扩展名调 用相应的文件解析模块对这些二进制文件进行解析,然后调用索引程序对分析好 的网页信息进行索引,建立索引数据库,最后供用户搜索。

在互联网上,HTML 文件是最常见的文档格式。Htmlparser 是一个专门用来解 析 HTML 文档的著名 Java 开源组件。Htmlparser 解析 HTML 的速度快,鲁棒性高。 甚至可以说,Htmlparser 是目前最好的 HTML 分析工具。另外,Htmlparser 结构 设计精良,便于扩展。

## 4.3网页结构解析模块

网页结构解析模块主要针对爬虫模块获取的网页HTML源码进行分析提取出网页的详细信息，包括网页正文、网页标题和网页中含有的超链接。

## 4.3.1 网页正文解析

网页的解析方法有很多中，基于模版适配提取，这种方法针对特定网站网页进行解析效果很好，但不具有通用性，智能新闻采集中的网页解析部分需要对任意网站都具有适用性，本文第三章提出的基于文本特征挖掘的网页文本提取方法可以适用于任意的网页，在本章智能Web新闻采集系统设计时讲采用本方法进行网页文本解析，提取出网页对应的正文信息和标题。

具体步骤：

1. 通过爬虫获取对应URL网页源码
2. 解析网页源码的编码方式，统一转化为UTF-8格式
3. 运用本文提出的基于文本标签特挖掘的网页正文提取方法提取网页的正文簇标签信息
4. 运用本文出的非新闻网页自识别方法进行网页判断，如果判断为非新闻网页，则不提取该网页的正文簇对应的标签文本信息，网页正文信息置空，如果是新闻网页，则根据正文簇标签ID组合各标签文本，形成网页的正文信息。
5. 返回网页正文信息。

通过正文解析模块，获取网页的正文信息，爬虫模块获取到网页正文后，根据是否为空判断其是否是新闻网页，如是新闻网页，这对正文进行存储。爬虫策略模块将对正文信息进行分词处理，确定URL优先级（策略问题）。

## 4.3.2 网页超链接解析

网络爬虫根据其爬虫策略不断的构造请求获取待爬取URL队列中网页信息，待爬取URL队列中的元素是从已获取网页HTML源码中提取出来的。目前比较方便的提取方式是通过DOM树解析方式获取HTML源码中的URL，但将HTML转化为DOM耗时，而且对不规则的HTML页面信息适用性差，本节将采用字符串解析的方式提取网页源码中的URL列表，并根据规则将提取出的URL转化为标准格式，在一个网页信息中存在各种噪声超链接，比如邮件网址、javascript网址等，在URL提取过程中需要将其过滤。其大致过程如下：

1. 解析网页HTML源码，获取网页内所有的URL组成的列表，并传入该网页的URL，并获取其网站域名
2. 遍历URL列表
3. 判断是否为噪声URL，若是则回退第3步
4. 判断URL是否是站内链接，若不是，则将改URL加入标准化后的URL列表，回退到第3步
5. 判断URL是否完整，若是，则将该URL加入标准化后的URL列表，回退到第3步
6. 根据该网页的URL和此站内不完整的URL的相对路径，将该站内不完整的URL补全为完整的URL，并加入到标准化后的URL列表，回退到第3步
7. 返回标准化后的URL列表

获取网页中的URL后，根据爬虫策略将URL加入待爬取网页的URL对列。。。。

## 4.3.3 非新闻网页识别

本系统时为了采集新闻网页信息，在获取爬虫获得网页信息之后，需要对网页进行判断，确定其是否是新闻网页，如果是新闻网页则进行网页正文存储，否则将不对网页内容进行存储；爬虫策略模块是根据网页内容和采集主题的相识度进行URL优先级排序，当判定该网页是非新闻网页后，则定其URL为最低优先级，否则根据相似度确定URL的优先级别。

在第三章中在基于网页文本标签挖掘的网页正文提取方法的基础上，提出了非新闻网页自识别方法，在对该方法进行验证后，其效果符合本新闻采集系统的要求，所以在系统设计时将不用基于URL特征的网页识别技术，而是采用本文提出的方法进行非新闻网页识别。

## 4.4 数据存储

数据采集只是网络舆情监控、搜索引擎和热点推送的基础工作，为了方便后续的数据分析工作，在数据采集过程中将对网页信息进行预处理后，不仅将其内容保存在本地磁盘，另外还将网页内容分词后的关键字列表和对应关键字权重信息存储到数据库中。

## 4.5本章小节

本章主要介绍智能Web新闻采集系统设计，首先详细讲解了新闻采集系统中爬虫模块的爬虫策略、初始种子选择和相应的分词方法，其次讲解了获取网页源码后系统如何对其进行解析及新闻网页的判断，网页解析过程将采用的本文提出的正文提取方法和非新闻网页识别方法。

# 智能Web新闻采集系统实现及总结

5.1系统框架

参考：基于标签路径特征的Web新闻内容抽取研究－吴共庆）

5.2系统运行效果评估