分 类 号 TP391 密级

主题网络爬虫关键技术研究

研 究 生 姓 名： 马 进

指导教师姓名、职称 ： 朱艳辉 教授

学 科 专 业 ： 计算机技术

研 究 方 向 ： 智能信息处理

湖 南 工 业 大 学

二〇一八年 六 月 二 日

分 类 号 TP391　　密级

主题网络爬虫关键技术研究

Research on key technology of subject network crawler

研 究 生 姓 名：　马 进

指导教师姓名、职称：　朱艳辉 教授

学 科 专 业：　计算机技术

研 究 方 向：　智能信息处理

论文答辩日期 2018.06.02　答辩委员会主席

湖 南 工 业 大 学

二〇一八 年 六 月 二 日

湖南工业大学学位论文原创性声明

本人郑重声明：所呈交的学位论文，是本人在导师的指导下，独立进行研究工作所取得的成果。除文中已经注明引用的内容外，本论文不含任何其他个人或集体已经发表或撰写过的作品成果。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标明。本人完全意识到本声明的法律结果由本人承担。

作者签名：马进　　日期：2018年 6月 2 日

湖南工业大学论文版权使用授权书

本人了解湖南工业大学有关保留、使用学位论文的规定，即：学校有权保留学位论文，允许学位论文被查阅和借阅；学校可以公布学位论文的全部或部分内容，可以采用复印、缩印或其他手段保存学位论文；学校可根据国家或湖南省有关部门规定送交学位论文。

作者签名：马进 导师签名：　日期：2018 年 6月 2 日

摘 要

随着互联网的迅速发展，Web 的信息量越来越大，人们往往通过搜索引擎去从互联网上搜索想要的信息，如：百度，谷歌，搜狗等。这类搜索引擎称之为通用搜索引擎，其为所有的用户提供用户想要的所有信息。随着互联网上的信息量越来越大，用户搜索出来的信息可能与自己想要的信息大相径庭。对于这种问题，就需要更加专业的、面向特定领域的搜索引擎来解决。

主题网络爬虫是垂直搜索引擎的关键部分，本文主要是对主题网络爬虫中的关键技术进行研究。主要研究内容如下：

(1) 主题内容的抽取是网页主题识别的重要步骤，本文结合网页内容分布特征以及主题内容的相关特征，设计了一种网页主题内容抽取方法。此方法首先将网页解析成 dom 树结构，然后根据对网页进行去噪去除掉网页的噪音节点，最后根据主题内容在页面中的分布特征去进行抽取。

(2) 提出了一种基于实体链接的主题识别算法，识别网页的主题。将基于知识库的实体链接方法运用于特征抽取，首先利用知识工厂提供的接口对原始语料进行分词并识别出语料中的实体，然后利用实体链接获取实体相关的信息，接着从实体信息中抽取出潜在的特征汇总到候选特征集合中，最后利用信息增益的方式从候选特征集合中挑选出最终的特征集合。最终利用抽取出的特征集合训练朴素贝叶斯分类器对网页主题进行识别。实验表明该方法提高了主题网页识别的准确率。

(3) 提出了一种改进的基于 Best-First 算法的主题搜索策略。主题搜索策略是指导主题网络爬虫抓取网页的关键，本文采用改进的基于 Best-First 算法的主题搜索策略。该策略主要思路是首先从待抓取链接列表中挑选出价值最大的链接进行抓取，然后从抓取到的网页中抽取链接，对这些链接的价值进行评估，如果链接价值小于设定的阈值则丢弃，反之则将其放入按照链接价值排序的待抓取队列中，循环此过程直到抓取深度到达预设值或者待抓取队列为空则停止。

关键词：主题网络爬虫，实体链接，Best-First 算法，主题搜索策略

ABSTRACT

With the rapid development of Internet, the amount of information in Web is increasing.People often use search engines to search the Internet for desired information, such as: Baidu, Google, Sogou, etc.This kind of search engine is called a general search engine, which provides all users with all the information they want.With the increasing amount of information on the Internet, the information searched by users may be different from the information they want.For this kind of problem, we need a more professional, search engine for specific areas to solve.The topic web crawler is a key part of the vertical search engine. This article mainly studies the key technologies in the topic web crawler.

This paper research content is as follows:

(1) The extraction of topic content is an important step in the topic recognition of a web page. This paper, based on the distribution characteristics of the web content and the related features of the topic content, designs a method for extracting web page subject content.This method first parses the webpage into a dom tree structure, then removes the noise nodes of the webpage according to the denoising of the webpage, and finally extracts according to the distribution characteristics of the theme content in the page.

(2) A topic recognition algorithm based on entity link is proposed to identify the theme of the webpage.The entity link method based on the knowledge base is applied to feature extraction. Firstly, the interface provided by the knowledge factory is used to segment the original corpora and identify entities in the corpora. Then entity links are used to obtain entity-related information. Then the potential features are extracted from the entity information into candidate feature sets, and finally used. The information gain approach picks the final feature set from the set of candidate features. Finally, the naive Bayesian classifier is trained on the web page subject using the extracted feature set.Experiments show that this method improves the accuracy of topic page recognition.

(3) An improved topic search strategy based on Best-First algorithm is proposed.Topic search strategy is the key to guide the theme web crawler to crawl web pages. This paper adopts topic search strategy based on Best-First

algorithm.The main idea of this strategy is to first select the most valuable link from the list of links to be crawled for crawling, then extract the links from the crawled pages, and then evaluate the value of these links if the link value is less than the setting. The threshold is discarded. Otherwise, it is placed in the queue to be fetched sorted according to the link value. This process is repeated until the crawl depth reaches the preset value or the crawl queue is empty.

Key Words: Theme web crawler; entity link; Best-First algorithm; topic search strategy

目　录

第一章　绪　论

1.1　背景与意义

随着Internet的飞速发展，互联网信息呈指数增长。根据中国互联网络信息中心（CNNIC）发布的第40次《中国互联网络发展状况统计报告》[1]数据显示：“截至2017年6月，中国网民规模达到7.51亿，占全球网民总数的五分之一。互联网普及率为54.3%；中国网站数量为506万个，半年增长4.8%。”

大量的网站中包含着不计其数的网页，网页是信息的载体，人们一般通过百度、谷歌等通用搜索引擎去从互联网上获取想要的信息。然而，利用通用搜索引擎搜索出的信息，往往比较宽泛，难以满足特定人群的需求。在这种情况，面向特定专业的搜索引擎，即垂直搜索引擎应运而生。垂直搜索引擎针对的是一个特定的行业，是通用搜索引擎的细分，其将某一领域的网页信息进行整合，处理后再以某种形式返回给用户。垂直搜索是相对通用搜索引擎的信息量大、查询不准确、深度不够等提出来的新的搜索引擎服务模式，通过针对某一特定领域、某一特定人群或某一特定需求提供的有一定价值的信息和相关服务。垂直搜索引擎专注于某一领域或专业，与通用搜索引擎相比，显得更加专注、具体及深入。

主题网络爬虫，又称聚焦爬虫是垂直搜索引擎的重要组成部分，所以对主题网络爬虫的研究具有重要的意义。主题网络爬虫是一个自动从互联网上抓取网页的程序，它根据预设的主题去访问互联网上与主题相关的链接，获取网页信息。与通用爬虫不同，主题网络爬虫并不追求大的覆盖，而将目标定为抓取与某一特定主题内容相关的网页，为面向主题的用户查询准备数据资源。通用网络爬虫从若干种子链接开始，先抓取种子链接的网页，然后从这些网页中抽取新的链接放入待抓取队列中，直到满足系统设定的抓取结束条件或者待抓取队列为空。主题网络爬虫的工作流程较为复杂，需要根据一定的网页分析算法过滤与主题无关的链接，保留有用的链接并将其放入待抓取的链接队列。然后，它将根据一定的搜索策略从待抓取链接队列中选择下一步要抓取的网页链接，并重复上述过程，直到达到系统的某一条件时停止。另外，所有被爬虫抓取的网页将会被系统存储到页面库，进行一定的分析、过滤，并建立索引，以便之后的查询和检索；对于主题网络爬虫来说，这一过程所得到的分析结果还可能对以后的抓取过程给出反馈和指导。

1.2 主题网络爬虫的国内外研究现状

1999 年，S.Chakrabani[2]第一次提出了聚焦爬虫这一概念，并设计并实现了 Focus Proiect 系统[3]。主题网络爬虫技术一经提出很快获得了广泛关注。国内外学者对其进行了深入的研究，不仅创新并设计了很多高效的主题识别算法，并且设计并实现了一些实用的主题爬虫系统。接下来，从理论与实现的系统两个方面介绍主题网络爬虫的国内外研究现状。

1.2.1 主题识别算法及主题搜索策略

P.DeBra[4]等人提出利用 Fish-Search 算法来作为爬虫的搜索策略，该算法假设主题相关页面逻辑上相接近来搜索主题相关的网页。Shark-Search 爬虫[5]是在 Fish-Search 算法的的基础上进行了改进，Fish-Search 算法是利用二值模型来评估主题相关性，而 Shark-Search 算法根据链接锚文本和网页主题相关内容计算出的相关性值为[0-1]内的值。该爬虫可以有效的找到相关信息，提高主题网络爬虫的召回率。Best-First 爬虫[6]，由 CHO J 等人在 1998 年提出，其主要内容是通过描写特定主题的关键词集合寻找种子链接，根据链接内容和链接对应网页的内容计算链接的相关性，依据计算出来的相关性值对待爬行队列进行排序，相关性值越大表示此链接具有较高的优先级，主题爬虫在抓取信息时优先抓取此链接。

Larry Page 和 Sergey Brin[7]提出了 PageRank 算法且运用于 Google 搜索引擎，Google 搜索引擎之所以成为全世界最流行的搜索引擎，除了高性能和高易用性以外，一个决定性的因素是它优秀的搜索结果。而它搜索结果的高质量来源于 PageRank 算法，该算法是一个精密的对网页文件重要度进行排序的算法。PageRank 算法的基本原理是，一个网页的重要程度依赖于它的入链，如果一个高重要度的网页文件链接到网页，那么根据 PageRank 的规则，此网页的等级越高。如此，根据 PageRank 算法，网页文件的重要度由与它链接的网页文件的重要度决定，而所链接的这些网页文件的重要度再由与它们链接的网页文件的重要度决定。因此，一个网页文件的 PageRank 由其它网页文件的 PageRank 总递归之和确定，总而言之，PageRank 的重要度由整个网络的链接结构决定。HITS 算法是有康奈尔大学的 Jon Kleinberg 博士于 1998 年首先提出的，该算法的目标就是通过一定的计算来得到针对某个检索的最有价值的网页。

Diligenti[8]利用“语境图”(Contex Graphs)构造分类器来指导爬虫爬行方向，用网页在语境图的层次来表示网页与主题网页的距离，距离越近的网页主题相关度越高，将优先被访问。

Johnson 等人提出基于 SVM 分类模型引导主题网络爬虫爬行[9]。Rennie 等人提出了面向机器学习的自适应算法引导主题网络爬虫爬行[10]，核心思想是利用学习算法引导爬虫以最小的代价穿越隧道到达相关页面。Gao 等人提出了聚焦协作爬行方法完成地理位置上的主题爬行[11]。Shokouhi 等人提出了一种名为 Gcrawler[12]的爬虫，其利用遗传算法估算最优路径。

陈军[13]提出了一种基于网页分块的的 Shark-Search 算法，该算法以块为基本单位计算链接的价值，能有效地去识别噪音链接块，该算法对包含较多噪音链接的网页具有较好的效果。

熊忠阳[14]等人提出一种基于信息自增益的主题爬虫搜索策略，该策略能使主题网络爬虫在爬行的过程中自动学习和更新。

1.2.2 主题爬虫系统

根据理论研究，国内外专家设计并实现了很多高效的主题爬虫系统。

(1) Scirus 系统。Scirus 系统[15]是由 Elsevier Science 和 FAST 合作开发的垂直搜索引擎，其为学生和科研工作者服务。它的主题爬虫进行信息收集时只爬行收录主题范围内的网页。此外，系统过滤从网络中搜索到的结果，只列出包含主题信息的搜索记录。作为互联网上最全面、综合性最强的科技文献门户网站之一，Scirus 系统曾多次被评为最佳专业搜索引擎。

(2) 美国国家数字科学图书馆 Collection Building Programe(CBP)系统。该系统主要面向于科学、数学在线数字图书。该系统的主要特点有两方面，一方面，系统只提供资源链接，而不对资源内容进行储存，如果用户需要查询网页资源，则需要通过系统提供的链接去互联网上获取资源。另一方面，其操作简单，用户只需要输入简单的查询信息，就能查询到相关度较高的链接。

(3) NEC 研究院的 CiteSeer 系统。CiteSeer（又名 ResearchIndex），是 NEC 研究院在自动引文索引(Autonomous Citation Indexing,ACI)机制的基础上建设的一个学术论文数字图书馆。这个引文索引系统提供了一种通过引文链接的检索文献的方式，目标是从多个方面促进学术文献的传播和反馈。CiteSeer 检索互联网上的 PostScript 和 PDF 两种格式的学术论文。目前，在 CiteSeer 数据库中可检索超过 500 万篇论文，这些论文涉及的内容主要是计算机领域。这个系统能够在网上提供完全免费的服务（包括下载 PostScript 或 PDF 格式的论文的全文）。该系统的主要功能有：①检索相关文献，浏览并下载论文全文；②查看某一具体文献的“引用”与“被引”情况；③查看某一篇论文的相关文献；④图表显示某一主题文献（或某一作者、机构所发表的文献）的时间分布。

(4) STIP 系统。该系统是中科院文献情报中心实施中科院文献信息共享系统的一个子课题，主要面向科技信息类资源。

(5) 南京大学的互联网数据采集系统(IDGS)。该系统采用模式匹配技术来实现自动搜索互联网上的中英文技术资料。

(6) 北大天网。该系统[16]采用一组关键词来表示一个主题，爬虫利用这组主题关键词按照策略从互联网中抓取数据，使其可以尽可能快且尽可能全面地抓取到与某主题相关的信息资源。

(6) 主题信息采集系统 Gsearch。由周鑫等设计并实现。该系统利用相似粗糙集和模糊认知图方法进行主题相关性判别，提供了主题信息的采集、存档、分析和检索功能。Gsearch 系统[17]在企业决策支持、行业市场分析等领域有着广泛的引用前景。

1.3 本文的研究内容

本文在通用网络爬虫的基础上，通过引入网页主题内容的提取以及基于实体链接的主题识别算法去识别主题网页，然后使用改进基于 Best-First 算法的主题搜索策略去指导主题网络爬虫从互联网上抓取主题相关的网页。

本文的研究内容如下：

(1) 集合网页内容分布特征以及主题内容的相关特征，设计了一种网页主题内容抽取方法。

(2) 在对主题网页的识别方面，采用基于实体链接的主题识别算法来识别主题网页。

(3) 在搜索策略上，采用改进的基于 Best-First 算法的主题搜索策略来指导主题网络爬虫抓取主题相关的网页。

本文共分为六章，篇节安排如下：

第一章，绪论。介绍了研究的背景与意义，主题网络爬虫的国内外研究现状，以及本文研究内容和篇章结构。

第二章，主要介绍了爬虫的体系结构。通过介绍通用网络爬虫和主题网络爬虫的体系结构来阐述主题网络爬虫与通用网络爬虫的区别。

第三章，主要介绍了网页主题内容的抽取。先介绍了 HTML 结构，然后介绍了网页的解析以及如何对网页进行去噪处理，最后阐述了如何抽取网页的主题内容以及分词的相关内容。

第四章，详细介绍了基于实体链接的主题识别算法。首先介绍了实体链接，然后阐述了如何将其使用到特征提取中，进而来提高主题识别算法的准确率。

第五章，提出了改进的基于 Best-First 算法的主题搜索策略。首先，介绍了通用网络爬虫的搜索策略以及相关算法，然后，介绍了两种主要的主题搜索策略以及相关的比较有代表性的算法，最后，详细阐述了本文所研究的改进的基于 Best-First 算法的主题搜索策略。

第六章，对论文工作进行总结与展望。

第二章　主题网络爬虫的体系结构

2.1　组成模块

2.1.1　基本组成

图 2-1 是主题网络爬虫的体系结构图

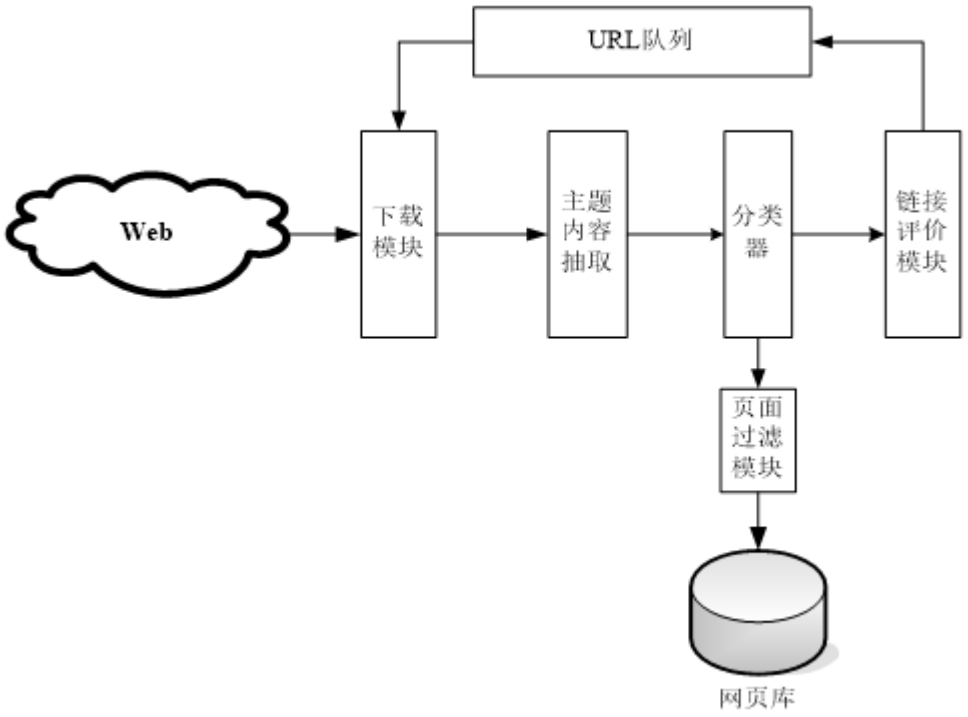

图 2-1　主题网络爬虫的体系结构

如图所示主题网络爬虫分为 4 个部分：下载、主题内容抽取、主题网页识别和链接评价模块。

(1) 下载模块。

对于网络爬虫来说，下载网页始终是其主要的工作。在下载模块中有一个链接调度模块，该模块根据搜索策略从待抓取链接队列中获取链接并将链接分发到各个下载线程。下载模块需要考虑到多方面的因素。在使用多线程下载网页的网络爬虫系统中，每个网页的下载都会启动一个线程，随着并发量的增加，会大量消耗系统资源，所以在多线程下载中，必须要考虑线程资源的调度。另外，有些链接无法访问到，如果抓取的过程中一直等待的会造成资源的浪费且影响爬虫系统的性能，所以必须要设定超时机制，舍弃掉等待时间过长的网页，节约系统资源并且提高爬虫性能。

(2) 主题内容抽取。

主题内容抽取在通用爬虫中并不多见，通常只是找出网页包含的链接，有些爬虫甚至不将网页表示成树结构而直接用模式匹配的方法得到链接。主题网络网络爬虫则需要细致地分析网页。主题内容抽取对网页的后续分析影响很大，消噪是主题内容抽取过程的重要环节。由于噪音内容的主题无关性，噪音内容会导致各个类别的特征不够明显：待分类网页中的噪音内容则会导致该网页类别不明确，影响网页主题识别的效果。除了消噪，主题内容抽取还包括中文分词、停用词删除，有些更加深入的分析甚至还包括自然语言的理解，如词法、句法分析等。

(3) 主题网页识别。

主题网页的识别就是判断抓取的网页是否与主题相关。从语义上看，一个主题可以一个概念、一个词语、一个短语，一个段落或一篇文章。主题概念的范围可大可小，可以非常抽象，但此时它的含义非常模糊；它的范围也可以非常具体，而此时它的意义却非常明确。主题选择是主题信息提取的基础。网页主题可以用很多其他表示方法，有些用关键词集合，有些用许多具有代表性的网页，有些用自然语言里的概念。主题用类目体系中的某些类别表示是一种简便可行的主题表示方法。在本文中，主题即预设的某一类信息资源的统称。主题选择是主题信息抽取的第一步，网页的主题由一组主题相关的特征来表示。

在本文使用文本分类技术来识别网页主题。其过程是：首先选定主题，然后准备主题相关的训练集，用特征向量表示网页，最后，利用分类算法对其进行分类。

(4) 链接主题相关性评价及抽取

首先去除掉明显的广告链接，然后将相对链接转换为绝对链接，最后评估链接的主题相关性并将其放入待抓取队列中。链接的主题相关性的计算主题要考虑父页面和链接锚文本的主题相关性。

2.1.2 基本流程

爬虫的基本流程可以分成下载过程和网页分析过程两个过程。下载过程主要的任务是从待抓取的链接队列中获取链接然后从互联网上下载网页，网页分析主要包括网页主题内容的抽取和主题网页的识别两个步骤。

(1) 下载过程

step1 调度模块从待提取链接队列中得到链接, 然后启动相应数量的下载线程。

step2 每个下载线程建立会话。

step3 建立连接然后下载网页。

step4 讲网页存储到本地, 然后再次获取待下载链接，并转到 step3，如果已经没有待下载链接则线程退出。

(2) 网页分析过程

step1 网页预处理模块先将原始网页构建成 dom 树。

step2 抽取出网页中所有锚文本及文本节点，分别存放到两个容器：anchors 和 texts 中。

step3 过滤掉无关节点。

step4 过滤噪音文本。

step5 根据网页主题内容的特征进一步抽取出网页的主题内容。

step6 对抽取出的网页主题内容进行分词处理。

step7 提取特征，用待分类的特征向量代表网页。

step8 预先用训练网页集合，训练基于朴素贝叶斯算法的分类器。将待分类向量用分类器分类，判断是否与主题相关。

step9 如果网页与主题相关，则将网页保存到网页库。

step10 从 anchors 中得到的所有锚节点，剔除一些链接，并评估链接的主题相关度。将新的链接及其主题相关度存到待抓取的链接队列中。

2.2 主题页面的分布特性

主题页面的分布往往符合四个特性：Hub/Authority 特性，Linkage/SiblingLocality 特性，站点的主题特性，隧道特性。

2.2.1 Hub/Authority 特性

美国康奈尔大学 Kleinberg 教授发现页面大体可以分成两种, 即中心页面和权威页面。中心页面往往含有许多链接，并且这些链接指向不同的主题。另外一种是权威页面，其倾向于同一主题，并且具有一定的权威性。Kleinberg 教授对一个页面引入 Hub 和 Authority 值体现上述特性[18]，并依据这种特性提出 HITS 算法。

2.2.2 Linkage/Sibling Locality 特性

Linkage 特性是指网页包含的链接所指向的网页的主题通常与该主题的主题相关。Sibling Locality 特性是指网页内同一区域的链接通常主题相关[19]。

2.2.3 站点的主题特性

一个站点往往包含一个或多个主题。往往相关主题的页面聚集在一起，而不同主题的页面团之间的链接较少。这主要是与人们在处理事务时使用分类的思维习惯有关。为了使用户能更加方便且高效地浏览网站上的资源，网站的设计者一般会将网站中同一主题的网页相互关联在一起。

2.2.4 隧道特性

主题页面分布还有一种特性，即站点上的各个主题页面团往往会通过一些主题无关链接连接在一起。这些链接像是横跨在主题页面团之间的隧道，这就是隧道特性。在抓取过程中，隧道会影响抓取效率。

2.3 搜索策略以及链接提取

搜索策略是网络爬虫爬虫从互联网上抓取网页的核心，其很大程度上决定了爬虫的效率。其中部分链接需要根据相关协议排除掉。链接的评分是搜索策略的关键，其决定了链接抓取的顺序。

2.3.1 robots 协议和相对链接的转换

2.3.1.1 robots.txt 文件和 META 标签

(1) robots.txt。往往网站的一些内容不希望被爬虫抓取。ROBOTS 开发界提供了两个解决方案：一个是 robots.txt，另一个是 META 标签。robots.txt 是存放于网站根目录下的文件名小写的一个纯文本文件，网站中不想被网络爬虫访问的部分可在该文件中申明。

“robots.txt”文件包含许多的记录, 每条记录的格式如下所示：

<field>:<optionalspace><value><optionalspace>

robots.txt 文件针对整个网站的，用来描述站点的爬虫的访问情况，而 META 标签则是主要用于单个具体的页面。

(2) META 标签中没有大小写之分，name=“Robots”表示作用于所有的网络爬虫，也可以针对某个具体网络爬虫写为 name=“BaiduSpider”。

2.3.1.2 相对链接的转换

相对 URL 有服务器相对 URL 和文档相对 URL。绝对 URL 的格式如下：

scheme: //server/path/resource

其中：scheme 指定资源所使用的协议，有 http，mailto，ftp 等协议。server 是指资源所在服务器的名称比如 www.baidu.com。path 是指到达资源的路径，比如 /18/0402/09。resource 通常是文件名比如：DECL75C900118017.html。它可能是单个二进制流的“简单文件”，也可能是“结构化文档”。定位资源的所有信息都包括在“绝对 URL”中。

相对 URL 相对于某一网页位置的目标链接。因为在现实环境中，网站服务器发生变更会引发链接错误，所以使用相对链接指向同一服务器下的网页。当前网页位置一般可视为特定网页位置，或者用 base 标签定义，如 <base href="http://mobile.163.com"/>，那么该网页中所有的链接都是以“http://mobile.163.com”为前缀。

2.3.2 搜索策略概述

通用网络爬虫为了较高的覆盖率，一般采用图的广度优先策略去遍历互联网上的网页：主题网络爬虫需要搜索的内容只会针对特定的主题，而不需遍历整个网络，只需要选择主题相关的网页进行遍历。

主题网络爬虫通常采用“最好优先”原则从互联网上搜索网页。每次对“最有价值”的链接进行访问来高效地获取到更多与主题相关的网页。主题网络爬虫不同的搜索策略由链接的价值评价方法决定。链接往往包含在页面内容之中，所以一般父页面的价值高，其所包含的链接一般也具有较高价值，因此对评价链接价值往往要结合对网页内容的分析。

搜索策略是主题网络爬虫的核心，其指导主题网络爬虫在互联网上抓取网页，而链接的评分是搜索策略的关键。近来，搜素策略不仅仅考虑内容或者链接，对于主题网络爬虫来说，只有结合链接和内容信息才能更好的指导其从互联网中抓取网页。所以，对于待抓取的链接的评分可以从以下三个方面来考虑：

(1) 父页面的预测分值。这个因素体现了主题页面分布的 Linkage/Sibling 特性。如果父页面主题相关，那么可以认为该页面上的其它链接所指向的网页很有可能也是主题相关的网页，所以一个链接会以一定的权值继承父页面的主题相似度。而父页面已经下载，其主题相似度是页面内容与主题向量计算的结果，所以这部分结合了对网页的内容分析和链接分析来对链接的评分进行预测。

(2) 锚文本。链接锚文本一般是对链接所指向的网页的精简描述，如果网页是主题相关的，那么链接的锚文本会含有主题相关的关键字。

(3) 链接结构。对于同一的网站来说，相同主题的网页的链接结构会很相似。

2.4 本章小结

本章概述了主题网络爬虫的基本流程和组成部分，介绍了链接提取规则。最后介绍了网络爬虫搜索策略的概念。

第三章　网页主题内容抽取

3.1　HTML 简介

HTML 是 HyperText Markup Language 的缩写，中文称为超文本标记语言，它是标准通用标记语言下的一个应用，也是一种规范和标准。通过 web 技术来创建 HTML 文件，其本身本身是一种文本文件，通过不同的标记，可以告诉浏览器如何显示它的内容。

目前大部分网页都是由 HTML 编写。网页通过超链接链接在一起，进而形成一个紧密连接在一起的网络结构。

大致可将 HTML 标签分为三类:

(1) 对网页进行布局的标签。此类标签主要是用来对网页内容进行布局，一般一个网页会包含有不同的区域，不同的区域包含不同的内容信息。这些区域一般通过这些进行布局的标签划分出来，常用的标签有<table>、<tr>、<td>、<p>、 等。

(2) 描述信息显示特点的标签。这些标签主要是用于告诉浏览器标签中的内容如何显示，比如：字体加粗、斜体等。常用的此类标签有<b>、<I>、、<h1>、<h2>等。

(3) 包含超链接的标签：超链接用于连接各个页面能表示网页之间的关系。这类标签有<a>、<img>、<frame>等。

HTML 文档主要由头部(head)和主体(body)组成[20]

HTML 文档主要有以下两个部分：

(1) 头部。这部分主要是对网页所需要资源的描述。主要包含这些标签：标题(<title>)、元信息(<meta>)、样式文件链接(<link>)、样式标签(<style>)、脚本(<script>)。

(2) 主体。这部分是网页显示的主要内容。主要包含这些标签：容器标签( )、布局标签(<table><tr><td>)、图片标签(<img>)、超链接标签(<a>)、段落(<p>)标题<h1>、换行(
)等。标签由标签名和属性两部分组成，属性由属性名和属性值组成。任何 HTML 文件都是以<html>标签开始，以</html>标签结束的。

3.2　网页文件解析

HTML 文件一般用 dom 树表示。解释 HTML 文件的过程就是将字符流表示成 HTML 树[21]的过程。

如图 3-1 所示的 html 树。

<html>

```
<head>
<title>标题</title>
</head>
<body>
<table>
<tr>
<a href="http://www.baidu.com/">百度</a>
<a href="http://www.163.com/">网易</a>
</tr>
<tr>
<p>段落</p>
</tr>
</table>
<body>
</html>
```

上述 html 文件的树结构表示如图 3-1 所示:

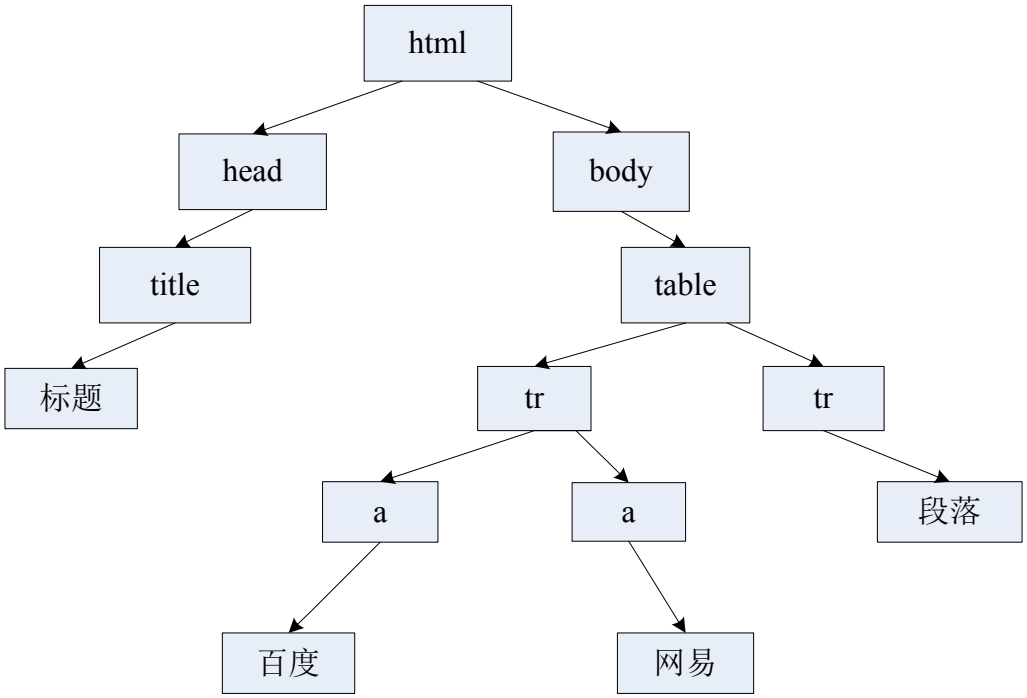

图3-1　html文件的树结构

目前，有很多构造标签树的工具，如：htmlParser，W3C HTML lexical analyzer[22]等。

3.3　网页去噪

许多网页都包含与主题内容无关的内容。如图 3-2 所示的新闻网页截图的网页，可以认为正文块以外的部分都是噪音。

从网页的截图可以看出，网页除了正文块外，其余部分由广告、导航链接、搜索服务等组成。

在主题搜索领域，大量的噪音内容会导致主题漂移。在提取主题相关内容时，如果将噪音内容作为主题相关的内容，影响对网页主题的识别。

用统计的方法去噪[23]的流程如下。

图3-2 网页的正文块

(1) 删除噪音块：网页去噪的基本方法是利用各种通用的特征来区分有效的正文和页眉、页脚、广告等其他信息。

(2) 划分段落。可以把 HTML 页面划分成多个段落(Paragraph)。简单的实现方法是，根据<td><p>
 <table>这些标签来划分段落。

(3) 评估段落。每个段落内的文字因其视觉上或者是主题上对文档的贡献程度而具有不同的权重。选取分值最大段落作为正文块。

去噪流程如图 3-3 所示。

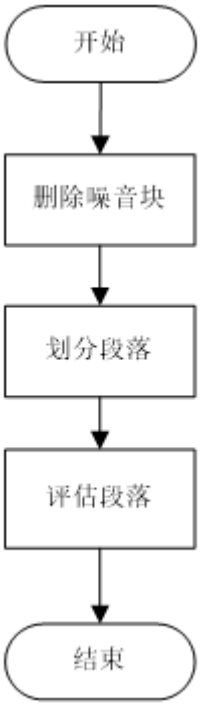

图3-3　去噪流程

3.4　主题内容的抽取

网页主题内容的抽取是网页主题识别的第一步，主题内容抽取的准确率直接影响到网页主题识别的准确率。刘军等[24]提出构建文档对象模型 DOM 树，然后添加显示语义等属性来解决 HTML 文档半结构化的不足，并提出一种聚类规则来对其进行分块，最后提取出主题信息。在此基础上，通过对大量网页的分析发现，除了正文的内容块，网页的标题(title 标签的内容)及 meta 标签的属性 content 的内容往往也携带有网页的主题信息，所以抽取的时候这部分内容也应该被抽取。网页一般可分为索引型网页和主题型网页，网页主题的识别是针对于主题型网页。主题网页所包含的链接文本往往是广告等与主题无关的文本，所以，应该先去除掉。然后还应去除掉网页版本信息等与主题无关的噪音文本。综上所述，网页主题内容的抽取算法如下：

Step1 首先选取 p，li，td，h1 作为分块节点

Step2 去除网页 a 标签及其内容

Step3 获取 title 标签中的文本，然后删除 title 标签及其内容

Step4 获取 meta 标签的属性 content 的值，然后删除 meta 标签

Step5 对 p 标签进行与Step3 相同的操作

Step6 对 li 标签进行与Step3 相同的操作

Step7 对于 td 标签进行与Step3 相同的操作

Step8 对于 h1 标签进行与Step3 相同的操作

取“，。？；、,.?”标点符号作为分块节点的特征，如果节点中的文本包含这些特征则认为内容是主题相关，就对其进行抽取。

3.5 文本分词

相比于英文网页，中文网页的处理往往有较大的不同。在中文里面，词是最小的有意义的语言成分，但是没有比较明显的标记去区分词语。所以，在处理中文文本之前都需要将文本切分成一个个单独的词语。

当前主流的分词算法切分精度差别不是特别大，但是切分时间差别较大，有的达到 20MB/s，比如利用 Unicode 编码进行优化的基于双数组 trie 树的分词算法[25]; 由于实现方面的缺陷, 有的算法的速度却只能达到 100K/s，如基于隐马模型的 ICTCLAS[26] 分词方法。目前分词算法主要分为两类：基于规则的和基于语料库的。

3.6 实验分析

3.6.1 实验环境

实验设备：一台 PC 机，CPU：Intel(R) Core(TM)i5，内存：8GB，硬盘：500GB

操作系统：Windows 10 专业版。

编程语言：JAVA。

开发环境：MyEclipse2014，JDK1.7。

3.6.2 实验结果与分析

以新浪网的一篇文章为例，对其进行主题内容抽取。原始页面截图内容如图 3-2 所示。主题内容抽取的结果如图 3-4 所示。

结果表明本小节所提出的网页主题内容的抽取能取得较好的效果。

中国国防部长常万全为何1年多4次赴此地
原标题：国防部长常万全为何1年多4赴此地？
近日，常万全以国务委员兼国防部长的身份，到云南调研边防工作。
据新华社报道，常万全强调，新时代边防的内涵在丰富，职能在拓展，要求在提高，必须认真贯彻党的十九大关于“建设强大稳固的现代边海空防”重大决策部署，创新工作思路，强化责任担当，不断开创边防工作新局面。要妥善应对新挑战，努力把我们的制度优势、经济优势、军事优势充分发挥出来，体现到强边固防的实际成效上。
从央视画面中看到，常万全此行来到了与缅甸掸邦接壤的云南省德宏州，视察了边防部队，探望当地群众，并身着迷彩服到边境察看。期间，他还在德宏州边防委员会联防中心视察座谈。
“政事儿”（微信ID：xjbzse）注意到，公开报道中，这是一年半以来，常万全第4次来到云南。两个月前，9月15日至16日，他还来到云南西双版纳州勐腊县、磨憨口岸，老挝琅南塔省、乌多姆赛省、磨丁口岸等地，与老挝国防部长占沙蒙举行边境高层会晤。
会晤期间，常万全与占沙蒙共同祭扫了位于老挝纳莫县的中国援老烈士陵园，种下了中老友谊树，参观了双方边防连队和勐腊县第一小学，观摩了两军边防部队联合巡逻，并发表了联合新闻公报。南部战区副司令员兼参谋长陈照海，南部战区空军政委徐西盛等陪同访问。
公报中称，双方一致同意密切两军高层交往，健全合作机制，深化务实合作，加强多边协调；继续开展边防友好交流合作，完善交流机制，丰富合作内容，定期组织联合巡逻、联合反恐等行动，共同维护好边境和平稳定。
今年2月，常万全到云南德宏州调研边防工作。
调研期间，他来到边防部队、边境检查站、爱国主义教育基地和边境村寨、国门小学，了解边境管控、边防基础设施、国防教育和边民生产生活等情况，与军地各级进行座谈交流，听取加强党政军警民合力强边固防的意见建议。
“政事儿”（微信ID：xjbzse）注意到，此次调研，常万全给边防战线作出指示：充分认清我国安全面临的复杂严峻形势。
他指出，边防战线的同志必须深刻领会习主席关于加强边海防工作的一系列重要指示，充分认清我国安全面临的复杂严峻形势，认清建设稳固边防在党和国家工作全局中的重大意义，认清党政军警民在合力强边固防中肩负的使命责任。
对于云南省的特殊地域位置，常万全强调，云南地处祖国西南边陲，区位优势独特，在国家安全和发展战略全局中具有特殊重要的地位作用。军地各级在稳边控边的工作指导中，要把治边优先的定位进一步摆到位，充分发挥党的绝对领导、政府的统筹协调、军队的中流砥柱、警方的治理监管、群众的重要依托作用。
2016年6月，常万全到云南普洱、临沧等地调研边防工作。
他到了边防一线哨所、边防检查站、口岸通道等地，看望边防官兵，了解战备、执勤、训练和生活情况；来到乡村和学校走访慰问当地群众。

图3-4　网页主题内容抽取结果

3.7　本章小结

本章先介绍了网页处理中的各种预处理过程，包括 HTML 解析、网页消噪。为了提取网页中的主要内容，有必要剔除原始网页的噪音。然后根据网页的主题内容的特征，提取出网页的主题内容。

第四章　基于实体链接的主题识别算法

4.1　实体链接简介

在介绍实体链接之前，首先，需要了解实体的概念。实体是存在于世界上的某一个对象或者对象的集合，实体一般用属性来描述。其中，一般通过实体表述(Mention)来代表一个实体，实体表述是文本中对该实体的引用[27]，实体表述一般可以分为三种形式[28]：名称表述(Name Mention)，名词或者名词短语表述(Nominal Mention)，以及代词表述(Pronoun Mention)。

实体链接就是把文本中的实体表述链接到知识库中的相应实体的过程[29]。在实体链接中所使用的知识库包括 Wikipedia、Freebase、YAGO、DBpedia 等[30]。目前复旦大学图数据管理实验室的知识工厂构建了知识库，并提供了比较全面的接口，本文所使用的就是知识工厂提供的知识库以及相关接口。在一段文本中，实体链接主要有两件事，一方便是识别出文本中的实体指称，另一方面就是将识别出的实体指称与知识库中的相应实体相关联。由于自然语言中普遍存在一词多义和别名现象，因此需要根据文本中实体表述的上下文信息去确定实体表述所指向的实体。所以，实体链接主要包含两项关键技术：实体识别、实体消歧。实体识别[31]是指识别一个文本中的实体表述，这个实体表述可能是指向实体的词或者短语。实体消歧[32]是指给定实体指称及其所在上下文、候选实体，判断其在当前上下文中所指向实体的过程。

例如：“[复旦大学]是国内顶尖的重点大学”。这句话中“[]”内的字符串就是实体表述。实体链接就是将这句话中的这个实体表述链接到它们在知识库中对应的实体上。图 4-1 就是将实体表述链接到复旦大学知识工厂的知识库中的实体信息。

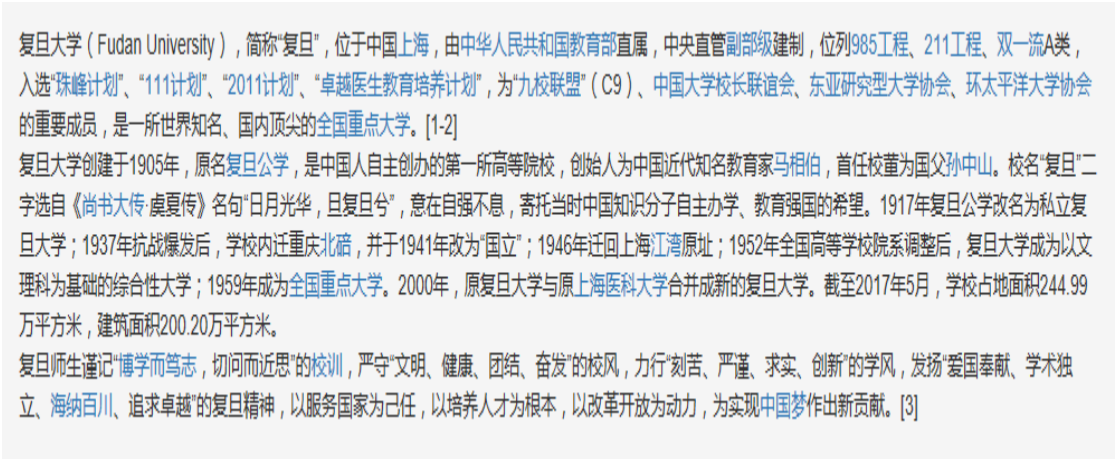
复旦大学（Fudan University），简称“复旦”，位于中国上海，由中华人民共和国教育部直属，中央直管副部级建制，位列985工程、211工程、双一流A类，入选“珠峰计划”、“111计划”、“2011计划”、“卓越医生教育培养计划”，为“九校联盟”（C9）、中国大学校长联谊会、东亚研究型大学协会、环太平洋大学协会的重要成员，是一所世界知名、国内顶尖的全国重点大学。[1-2]
复旦大学创建于1905年，原名复旦公学，是中国人自主创办的第一所高等院校，创始人为中国近代知名教育家马相伯，首任校董为国父孙中山。校名“复旦”二字选自《尚书大传·虞夏传》名句“日月光华，旦复旦兮”，意在自强不息，寄托当时中国知识分子自主办学、教育强国的希望。1917年复旦公学改名为私立复旦大学；1937年抗战爆发后，学校内迁重庆北碚，并于1941年改为“国立”；1946年迁回上海江湾原址；1952年全国高等学校院系调整后，复旦大学成为以文理科为基础的综合性大学；1959年成为全国重点大学。2000年，原复旦大学与原上海医科大学合并成新的复旦大学。截至2017年5月，学校占地面积244.99万平方米，建筑面积200.20万平方米。
复旦师生谨记“博学而笃志，切问而近思”的校训，严守“文明、健康、团结、奋发”的校风，力行“刻苦、严谨、求实、创新”的学风，发扬“爱国奉献、学术独立、海纳百川、追求卓越”的复旦精神，以服务国家为己任，以培养人才为根本，以改革开放为动力，为实现中国梦作出新贡献。[3]

图4-1　实体信息

4.2 CN-DBpedia

CN-DBpedia[33]是由复旦大学知识工场实验室（Knowledge Works）研发并维护的大规模通用领域结构化百科，其前身是复旦GDM中文知识图谱，是国内最早推出的也是目前最大规模的开放百科中文知识图谱，涵盖数千万实体和数亿级的关系。CN-DBpedia以通用百科知识沉淀为主线，以垂直纵深领域图谱积累为支线，致力于为机器语义理解提供了丰富的背景知识，为实现机器语言认知提供必要支撑。CN-DBpedia已经从百科领域延伸至法律、工商、金融、文娱、科技、军事、教育、医疗等十多个垂直领域，为各类行业智能化应用提供支撑性知识服务，目前已有近百家单位在使用。CN-DBpedia具有体量巨大、质量精良、实时更新、丰富的API服务等特色。CN-DBpedia已经成为业界开放中文知识图谱的首选。本文主要用CN-DBpedia作为知识库，并利用知识工厂提供的相关接口识别文本中的实体，然后进行再利用其提供的接口进行实体链接。

如图4-2文本:

北空某导弹团今天传出喜讯，营参谋长谭正提出的兵器改进方案，使困扰该团和兵器设计厂家三载的兵器重大隐患迎刃而解，赢得了官兵和有关专家的赞扬。

图4-2 原始文本

调用知识工厂的接口，获取到如图4-3的结果:

{"cuts": ["北空", "某", "导弹", "团", "今天", "传出", "喜讯", " , ", "营", "参谋长", "谭正", "提出", "的", "兵器", "改进", "方案", " , ", "使", "困扰", "该团", "和", "兵器", "设计", "厂家", "三载", "的", "兵器", "重大", "隐患", "迎刃而解", " , ", "赢得", "了", "官兵", "和", "有关", "专家", "的", "赞扬", "。"], "entities": [[[14, 17], "参谋长"]]}

图4-3 实体识别并分词后的文本

其中cuts字段是对文本的词集合，entities字段是识别出的文本中的实体。

然后调用知识工厂提供的获取知识库中的实体详情的接口，获取的实体信息如图4-4所示:

{"status": "ok", "ret": [["拼音", "cān móu zhǎng"], ["中文名", "参谋长"], ["亦称", "首长"], ["外文名称", "chief of staff"], ["CATEGORY_ZH", "组织机构"], ["DESC", "参谋长是各级部队军事指挥部门的首长，协助该部队的军事主官进行指挥。 军团以上包括旅、师、集团军、军区等各级部队都有参谋长。在团以下单位，只有作战参谋等职，而没有参谋长一职。"]]}

图4-4 实体详情

其中，status表示接口调用状态，ret字段表示获取的结果，其中包含实体的详细信息：拼音、中文名、亦称、外文名称、CATEGORY_ZH、DESC。

上述过程就是识别文本中的实体并链接到知识库中获取实体信息的过程，也即实体链接的过程。

4.3 基于实体链接的特征抽取

特征抽取对于主题识别来说，是特别重要的一步，抽取出的特征准确与否，决定着主题识别的准确率。传统的方式，由于分词的问题，无法很好的把某些主题特征抽取出来，所以本文将实体链接引入到特征抽取的过程中，以知识库为支撑，来更加准确的将主题特征抽取出来。

4.3.1 候选特征集合抽取

首先，从训练语料中抽取出候选特征集合[34]。

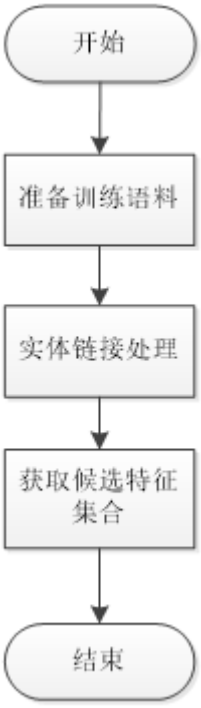

图4-5 候选特征集合抽取流程

如图 4-5 所示，具体步骤如下：

(1) 准备训练语料。准备若干主题相关的文本。

(2) 实体链接处理。使用知识工厂的接口将这些主题相关的文本逐句进行实体识别并分词，且获取实体信息从实体信息中抽取出更多的候选特征。

(3) 获取候选特征集合。将上一步获取到的词集合进行去重，去停用词处理，获取到候选特征集合。

从搜狗语料中选择与“军事”主题相关 50 篇的新闻文章作为选择候选特征的语料集合。首先，利用知识工程的接口对语料文章进行分词并识别语料文章的实体，由

于有知识库的支撑，所以能更好的将文章中的潜在的特征切分出来。对于识别出的实体进行实体链接，获取到实体的信息，如图 4-6 所示。

{"status": "ok", "ret": [["中文名", "防空导弹"], ["外文名称", "Air defense missile"], ["别名", "面空导弹"], ["别名", "地空导弹"], ["类别", "军事"], ["类别", "武器"], ["CATEGORY_ZH", "科学百科航空航天分类"], ["CATEGORY_ZH", "社会"], ["CATEGORY_ZH", "武器装备"], ["DESC", "防空导弹是指由地面、舰船或者潜艇发射，拦截空中目标的导弹，西方也称之为面空导弹。"]]}

图4-6　实体信息

上图所示，实体“防空导弹”的实体信息中，可将“别称”、“类别”及“CATEGORY_ZH”属性抽取出来合并到候选特征集合中。“DESC”属性加入到提取候选特征的语料集合中做进一步处理。

如上所述，基于实体链接的方式，由于有知识库的支撑，一方面，能较准确地将潜在的特征抽取出来，另一方面，由于抽取特征的训练语料是有限的，可以从实体信息中抽取出更多的特征。

图 4-7 为使用了实体链接所挑选出的候选特征集合实例。

防空
装甲旅
配套
朝鲜战争
发挥
六年
160
林肯
较少
伊朗政府
贺龙
必败
战斧
军
步营
300
座
晨曦
地铺
记
路线
抛锚
每艘
关键性
全长
9m317
双垂尾

图4-7　候选特征部分截图

4.3.2 常见特征抽取算法

目前，常用的特征选择算法如下：文档频率[35]、信息增益[36]、互信息[37]及词条的统计[38]等。

(1) 文档频率。

训练语料中包含某一词语的文档的条数就是该词语的文档频率。该方法的基本思想是：出现频率较低的词语往往携带很少的信息量，因而无法很好的将类别区分开来，因此可以删除较低频率的词语，这样既能降低特征维度又能提高分类的准确率。

(2) 信息增益(Information Gain)

信息增益(IG)是计算某一特征出现出现和不出现两种情况下，系统携带信息量的差值。对于文本分类而言，包含特征词 t 和不包含特征词 t 的文档频数差值代表了特征词 t 的 IG 值。IG 值采用如下的公式计算：

$$\begin{aligned} \mathrm{IG}(t) &= -\sum_{i=1}^{n} P(C_i)\log P(C_i) + P(t)\sum_{i=1}^{n} P(C_i \mid t)\log P(C_i \mid t) \\ &+ P(\bar{t})\sum_{i=1}^{n} P(C_i \mid \bar{t})\log P(C_i \mid \bar{t}) \end{aligned} \tag{4-1}$$

公式中，$P(C_i)$ 表示 C_i 类文档出现的概率，$P(t)$ 表示包含词 t 的文档概率，$P(C_i \mid t)$ 表示文档包含词 t 时属于 C_i 类的条件概率，$P(\bar{t})$ 表示不包含词语 t 的文档概率，$P(C_i \mid \bar{t})$ 表示文档不包含词语 t 时属于 C_i 类的条件概率，n 表示训练语料的类别数。

(3) CHI 统计

CHI 统计常常称为开方统计，用于检验两个变量是否独立。在两个变量是相互独立的前提下，将样本的实际观测值和理论值的偏离程度计算出来，表示为 CHI 值。CHI 值越大，两变量趋于相关；反之则两变量趋于独立。

特征词和文档类别的相关程度也可以用此方式来衡量。先预设词条与某一类别是独立的，以此为基础计算出的词条的 CHI 值越大则说明结果与假设偏差越大，则该词条与类别越相关。因此，该方法特征选择的过程即：计算每个词条与类别的 CHI 值，并从大到小排序，靠前的值则为特征。词语 t 对于类别 C_i 的 CHI 值计算公式如下：

$$CHI(t, C_i) = \frac{N \times (AD - CB)^2}{(A + C)(B + D)(A + B) + (C + D)} \tag{4-2}$$

公式中，N表示语料中的文档总数，A 表示包含词语 t 且属于 C_i 类的文档数，B 表示包含 t 但不属于 C_i 类的文档数，C 表示不包含 t 但属于 C_i 类的文档数，D 表示既不包含词语 t 也不属于 C_i 类也的文档数。

(4) 互信息(Mutual Information)

在信息论中，互信息(MI)表示两个事件发生相关联而提供的信息量。互信息量越大，相关性也就越大。

词语t和类别C_i的互信息计算公式如下：

$$MI(t, C_i) \approx \log \frac{A \times N}{(A + C)(A + B)} \tag{4-3}$$

公式(4-3)中变量的含义与公式(4-2)中一致。

实验表明，信息增益能比较有效的进行特征提取，所以本文采用信息增益的算法进行特征提取。

4.3.3 最终特征抽取

在候选特征集合的基础上，抽取最终的特征集合，具体步骤如图 4-8 所示。

图4-8 最终特征抽取流程

(1) 计算信息增益值。在训练语料中计算候选特征集合中每个特征的信息增益值。

(2) 获取特征。按照信息增益值递减排序，选择靠前的若干特征。

(3) 获取最终特征。将从实体信息中获取到特征加入到(2)步骤中获取的特征集合中，将这个特征集合作为最终的特征集合。

4.4 基于朴素贝叶斯算法的分类器

本文采用朴素贝叶斯算法构造分类器。

假设用特征集合来表示每个实例 A，而类 c 从某有限集合 C 中取值。现提供一训练实例集和一测试实例(a1，a2，…，am)。

需要分类的实例 A 的目标是获取实例(a1，a2，…，am)的类标记 $c(a)$。得到：

$$c(a)=\underset{c\in C}{\operatorname{argmax}}P(a_1,a_2,...,a_m\mid c)P(c) \tag{4-4}$$

现在要做的就是基于训练实例集估计式(4-4)中的两个概率值。

朴素贝叶斯分类器(naïve Bayes classifiers)[39]假定：在给定类标记时属性值之间是相互条件独立的。也就是说，联合概率正好是每个单独特征概率的乘积。具体的公式如下：

$$P(a_1,a_2,...,a_m\mid c)=\prod_{j=1}^{m}P(a_j\mid c) \tag{4-5}$$

代入式(4-4)中，可得朴素贝叶斯分类器的分类公式：

$$c(a)=\arg\max_{c\in C}\ P(c)\prod_{j=1}^{m}P(a_j\mid c) \tag{4-6}$$

式中，a_j 为 x 的第 j 个特征值、概率 p(c)和 $P(a_j\mid c)$ 可以通过计算训练实例集中不同类和特征值组合的出现频率来简单计算，具体的公式如下：

$$P(c)=\frac{\sum_{i=1}^{n}\delta(c_i,c)}{n} \tag{4-7}$$

$$P(a\mid c)=\frac{\sum_{i=1}^{n}\delta(a_{ij},a_j)\delta(c_i,c)}{\sum_{i=1}^{n}\delta(c_i,c)} \tag{4-8}$$

式中，n 为训练实例的个数、c_i 为第 i 个训练实例的类标记、a_{ij} 为第 i 个训练实例的第 j 个属性值，$\delta(c_i,c)$ 为一个二值函数，当 $c_i=c$ 时为 1，否则为 0。

显然，当出现零频率属性值的时候，这种方法会导致过低估计概率。更极端的情况会使得某个概率值为 0，进而导致由式(4-6)计算的整个量为 0。常常使用 Laplace 估计来进行平滑处理进而避免上述问题, 重写式(4-7)和式(4-8)得到:

$$P(c)=\frac{\sum_{i=1}^{n}\delta(c_i,c)+1}{n+n_c} \tag{4-9}$$

$$P(a\mid c)=\frac{\sum_{i=1}^{n}\delta(a_{ij},a_j)\delta(c_i,c)+1}{\sum_{i=1}^{n}\delta(c_i,c)+n_j} \tag{4-10}$$

式中，n_c 为类的个数、n_j 为训练实例第 j 个属性的取值个数。

基于实体链接的朴素贝叶斯分类器的整体框架如图 4-9 所示。

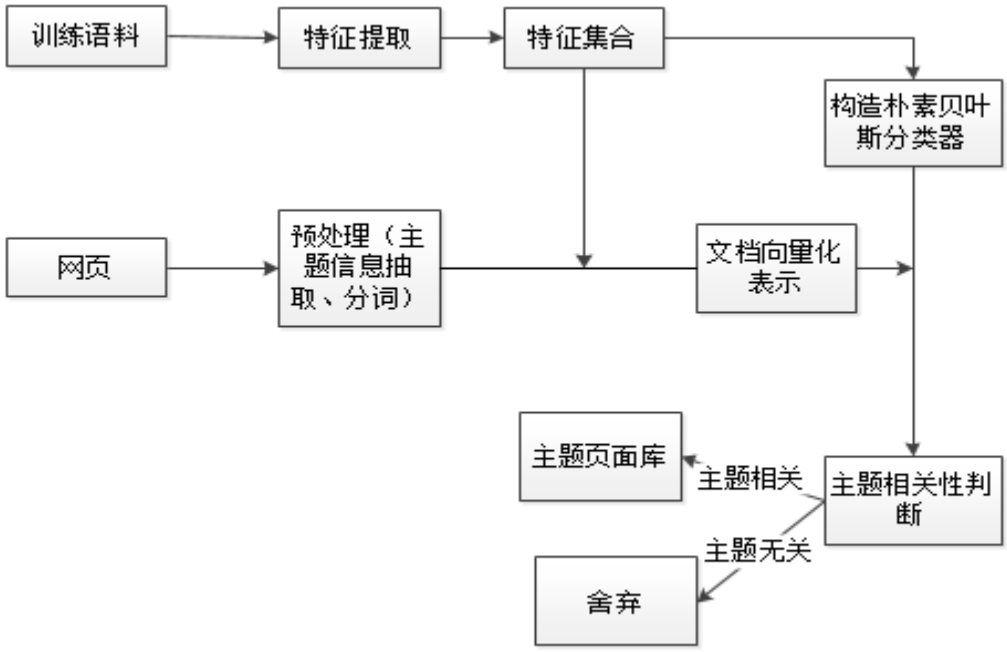

图 4-9　分类器构造及主题网页识别

其工作流程如下：

(1) 使用基于实体链接的特征抽取方法进行特征提取。

(2) 根据获取到的特征集合，构建朴素贝叶斯分类器，并对其进行训练。

(3) 对爬虫抓取到的网页进行预处理，包括：主题信息抽取、分词等预处理，然后将网页向量化表示。

(4) 利用分类器对向量化处理后的网页进行主题识别，如果网页属于主题类，则将网页保存到主题页面库，否则舍弃该页面。

4.5　实验分析

4.5.1　实验环境

本章所使用的实验环境与 3.6.1 小节相同。

4.5.2　实验结果与分析

本小节来通过实验验证本章提出的基于实体链接的朴素贝叶斯分类器在主题识别中的效果。

对于主题识别效果的评判，主要采用三个指标：准确率(P)、召回率(R)和 F 值[40]。准确率是表示准确识别出的主题相关的文本数量比例；召回率是准确识别出的主题相关的文本数量与训练集中所有主题相关的文本数量的比例；F 值是一个综合评价指标。

假定：在训练语料中，属于与主题相关且被判定为与主题相关的文本是 a，与主题无关但被判定为主题相关的文本数目是 b，属于与主题相关但未能被判定为与主题相关的文本数量是 c，那么三个评价指标的计算公式如下：

准确率： $P=\frac{a}{a+b}$

召回率： $R=\frac{a}{a+c}$

F 值： $F=\frac{2\times(P\times R)}{P+R}$

从搜狗新闻语料中选择军事（587 篇）及非军事（856 篇）的文章共计 1443 篇，作为训练语料。使用本章提出基于实体链接的方法去构建朴素贝叶斯分类器，进行实验。实验结果如下：

表4-1　实验结果

	识别出主题相关文本	识别正确的文本	准确率(P)	召回率(R)	F 值(F)
NB	613	516	84.2%	87.9%	86%
基于实体链接的 NB	627	552	80%	94%	90.9%

从实验结果来看，相比与传统的朴素贝叶斯分类器，引入实体链接技术对其进行改进能取得较好的效果。

4.6　本章小结

本章主要介绍了基于实体链接的朴素贝叶斯分类器。首先，介绍了实体链接的相关概念，接着介绍了本文所用的知识库 CN-DBpedia 以及相关接口。然后，重点阐述了基于实体链接的特征抽取方法，将实体链接技术运用于特征抽取中来更好的提取出主题相关的特征。接着，详细介绍了基于朴素贝叶斯算法的分类器的构造以及工作流程。最后，通过实验分析，证明本章提出的基于实体链接的主题识别算法能取得较好的效果。

第五章　改进的基于 Best-First 算法的主题搜索策略

搜索策略是爬虫预设的一种爬行方法，用于指导网络爬虫抓取互联网上的网页。本章在前人的基础上，提出了一种改进的基于 Best-First 算法的主题爬虫策略去指导主题爬虫去抓取网页。

5.1　通用搜索策略

互联网可以看成一个复杂而庞大的连通图[41]。参照图的遍历方法，通用网络爬虫一般有两种遍历策略，即广度优先策略和深度优先策略[42]。

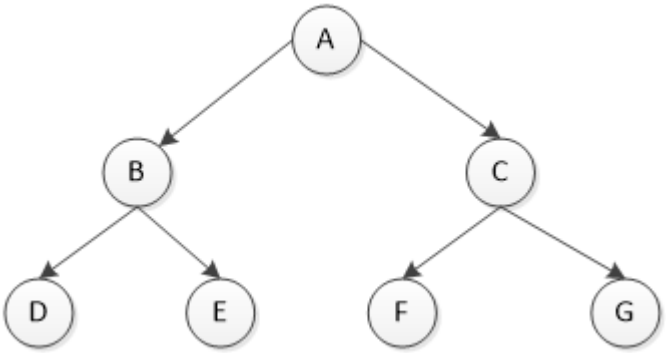

图 5-1　网络链接结构图

图 5-1 是个简单的网络链接结构图，其中网络中的网页由节点表示，而链接则由边来表示。深度优先搜索的思路是，爬虫沿着某一方向一直搜索，直到此方向没有可继续搜索的节点则换个方向继续进行。以图 5-1 为例，爬虫的搜索顺序为 ABDECFG。广度有限搜索策略的思路是由近及远一层层抓取搜索网络上的节点，同样以图 5-1 所代表的网络结构为例，其中，A 节点为第一层页面，B、C 节点为第二层页面，D、E、F、G 为第三层页面，所以按照广度优先的策略，爬虫的抓取网页的路径为 ABCDEFG。

理论上来说，这两种搜索策略都是可行的。但是，真实环境的互联网庞大且复杂，深度优先策略往往会陷入某一方向，而且，一般网页的层次越深价值越低，所以，通用网络爬虫一般采用广度优先策略。然而，通常通用网络爬虫为了追求覆盖率，所以抓取到的资源往往价值不高。

5.2 常用主题搜索策略

主题网络爬虫的目标是抓取与给定主题相关的网页，所以，需要预测链接内容的主题相关性，然后决定是否进行抓取。不同于通用搜索策略，主题搜索策略是在主题的指导下，对链接的价值进行评估。根据链接的主题相关性的大小，将链接插入到待抓取队列中，并选择主题相关度高的链接继续抓取[43]。这样主题网络爬虫能尽可能快且多的抓取到主题相关的网页，提高了抓取效率。

近年来，人们对主题搜索策略主要分为两类：一种是基于内容评价，另一种是基于链接结构评价[44]。

5.2.1 基于内容评价的搜索策略

该搜索策略主要是通过分析网页内容与所给主题的相关性来指导主题网络爬虫抓取网页。下面主要通过介绍如下算法来对这类搜索策略进行分析。

(1) Best-First 算法

最佳优先搜索(Best First Search)，是一种启发式搜索算法，它可以看做是广度优先搜索算法的一种改进。最佳优先搜索算法在广度优先搜索的基础上，用启发估价函数对将要被遍历到的点进行评价，然后选择代价小的进行遍历，直到找到目标节点或者遍历完所有点。用来做主题网络爬虫的搜索策略的基本思路是：对链接的价值进行评估，优先抓取价值高的链接所指向的网页。由于主题页面分布具有 Linkage Locality/Sibling 特性，所以如果父页面是与主题相关的页面，那么子页面一般也会继承父页面的相关性。因此，可以通过计算父页面的主题相关度来对评估链接的价值。

(2) Fish-Search 算法

De Bra 等于提出 Fish-Search 算法，该算法模拟鱼群觅食行为。该算法假设每个链接就是一条鱼，该链接的页面中的链接代表鱼的后代。如果鱼找到事务，即链接找到主题相关页面，那么将继续沿这个方向继续搜索。反之，如果鱼找不到食物，则后代变得虚弱，即此方向找不到主题相关页面。经过多次寻找，直到此线路再无其他链接则从其他方向重新开始搜索。

Fish-Search 算法根据用户提供的种子页面动态维护一个链接优先队列并设定初始搜索深度。如果当前链接找到主题相关网页，那么后代链接将继承当前链接的深度值，否则，后代链接深度值减 1。当某一个方向上链接深度值减为 0 时，舍弃该方向搜索。

(3) Shark-Search 算法

Fish-Search 算法使用二值模型来来判断鱼是否找到食物，这样无法精确的对链接优先队列进行排序。Hersovici 提出了 Shark-Search 算法来改进 Fish-Search 算法。其相关度一般在 0 和 1 之间取值[45]。

5.2.2 基于链接结构评价的搜索策略

互联网上的页面都不是孤立的，它们通过链接相互联系起来。根据互联网页面分布的 Hub 特性，权威值较高的页面往往被多个页面所指向。网页之间的这种链接结构关系能够一定程度上表征页面的重要性，对于预测链接价值有较大的帮助。基于链接结构评价的搜索策略中，PageRank 算法和 Hits 算法是最为基础、典型的两种算法。

(1) PageRank 算法。

定义“入度”和“出度”的概念，“入度”即指向网页的链接，“出度”即网页中指向其他页面的链接。PageRank 算法主要是依据入度和出度，对链接价值进行评估。

页面 A 的 PageRank 的计算公式如下:

$$\mathrm{PR}(A)=(1-d)+d\times\sum_{i-1}^{n}\frac{PR(T_i)}{C(T_i)} \tag{5-1}$$

在公式中： PR(A) 代表网页A的PageRank值。 $PR(T_i)$ 代表链接到A的网页 T_i 的PageRank值。 $C(T_i)$ 代表网页Ti的出度。d是阻尼系数，主要起调控作用，在 0 到 1 之间取值，一般设置d为 0.85。

(2) HITS 算法。

HITS 算法是一种基于超链接关系衡量网页重要程度的算法。根据页面分布的 Hub/Authority 特性，页面之间有一种相互加强的关系[46]：高质量 Authority 页面会被较多高质量 Hub 页面所指向，而高质量 Hub 页面将链接到较多高质量 Authority 页面上。

HITS 算法的基本流程为：

1) 网页 i 具有两个特征分值：中心度 h(i)和权威度 a(i)。初始情况下， h(i)=1，a(i)=1。

2) 每次通过迭代计算页面的中心度和权威度。

网页的中心度 h (i)为： $\mathrm{h}(i)=\sum a(i)$

网页的权威度 a (i)为： $a(i)=\sum h(i)$

标准化处理： $h(i)=h(i)/|h(i)|$， $a(i)=a(i)/|a(i)|$。其中，|h(i)|和|a(i)|分别代表了网页集合里的最大中心度和最大权威度。

3) 不断重复 2)的过程，计算上一轮迭代的权值和本轮迭代之后权值的差异，如果二者差异较小，则说明系统已趋于稳定。

5.3 基于 Best-First 算法的主题搜索策略

5.3.1 链接价值评估

Best-First 算法的基本思想是构建一个待抓取链接列表，然后从中选择最有价值的链接进行搜索。通常利用页面内容与主题的相似度来对页面的价值进行评估。利用向量空间模型来表示一个页面，将欧式距离和权重计算相结合的方法计算出页面与主题之间的相关性，其计算公式如下：

$$dist(q,p)=1-\sqrt{\frac{\sum_{i=i}^{n}(\omega_{iq}-\omega_{ip})^2}{\sum_{i=i}^{n}(\omega_{ip})^2}} \tag{5-2}$$

以上公式在欧式距离计算公式的基础上进行归一化并取反处理后的公式，其中，q，p 分别表示主题向量和页面特征向量，向量的维数都为 n，ω_{iq} 表示特征关键字 i 在页面 p 中权重，ω_{ip} 表示 k 在主题向量 q 中的权重，权重采用 TF-IDF 方式计算。

公式表示，当计算值最小为 0，表示网页向量与主题向量相似度最低，当计算值最大为 1，此时表示网页向量与主题向量相似度最高。

一个主题相关的网页上，并非所有的链接都是主题相关的；反之，一个主题不相关的网页上不一定所有的链接都是主题无关的[47]。所以链接的价值应该由两部分来评估，一部分是链接继承父页面的价值；另一部分则是链接的锚文本。

假设父页面的向量表示为 d_1，链接锚文本的向量表示为 d_2，主题向量表示为 q，通过前文的分析，得到基于内容的计算链接价值的公式：

$$v=\alpha\times dist(d_1,q)+\beta\times dist(d_2,q) \tag{5-3}$$

其中α，β为阈值调节参数，α+β=1。$dist(d_1,q)$表示父页面与主题的相似度，$dist(d_2,q)$表示链接锚文本与主题的相似度，主题相似度由公式(5-2)计算得出。

以上是基于对内容的分析，然而，对于一个站点来说，内容结构往往呈现一定的规律。这种规律性往往会体现在链接结构上，一个较大的站点一般会将站点中的网页分为：索引页，栏目索引页，内容页。索引页是站点的入口，包括站点所有的栏目以及部分网站重点内容页的链接。栏目索引页包含着同一主题内容的网页，而同一栏目下的网页的链接的结构是相似的。如表 5-1 所示。同一站点下的相同主题的页面的链接结构一般高度相似。所以将链接结构的因素加入到加入到链接价值计算中，来对基于 Best-First 算法的主题搜索策略进行改进。

表5-1 链接结构示例

网页类型	链接结构
索引页	http://www.sina.com.cn/
栏目索引页	http://mil.news.sina.com.cn/
内容页	http://mil.news.sina.com.cn/china/2018-04-13/doc-ifzcyxmu0751199.shtml http://mil.news.sina.com.cn/china/2018-04-14/doc-ifzcyxmu4097059.shtml

具体的做法是在公式(5-3)中加入调节因子 k，最终的基于内容与链接结构的链接价值计算公式如下：

$$v=\alpha\times(1+k)\times dist(d_1,q)+\beta\times dist(d_2,q) \tag{5-4}$$

调整因子 k 主要受父页面链接结构影响，所以如公式(5-4)所示，将调整因子加到父页面相关的计算部分。至于调整因子的计算主要考虑如下两个方面：

(1) 与父页面链接结构相似。和父页面链接具有相同的域名且其他部分结构也相同。

(2) 与父页面链接域名相同。和父页面链接具有相同的域名但是其他部分结构不同。

基于上述两个方面的考虑，调整因子 k 的取值如下表：

表5-2 调整因子k的取值

与父页面链接关系	k
与父页面链接结构相似	1
与父页面链接域名相同	1/2
其他	0

5.3.2 主题搜索策略

本文所使用的搜索策略，主要依赖于两个队列来指导主题网络爬虫来抓取网络上的网页资源。一个是待抓取链接队列，另一个是已抓取链接队列。其中，待抓取链接队列是本文搜索策略的关键。

待抓取链接队列示例：

[{depth=2,

url=http://tech.sina.com.cn/mobile/n/apple/2015-02-12/081510009651.shtml,

thematicCorrelationVal=0.8}]

待抓取链接队列结构如上所示，主要由三个元素组成：url，depth，thematicCorrelationVal。url 表示待抓取的链接，depth 表示链接深度，

thematicCorrelationVal 表示链接的计算出的主题相关度。待抓取链接队列是按照主题相关度递减排序的有序队列。

搜索策略的基本思路是：先从种子链接开始抓取，抽取页面上的链接，计算链接主题相关度，将链接按主题相关度大小插入待抓取链接队列中。详情如图 5-2 所示。

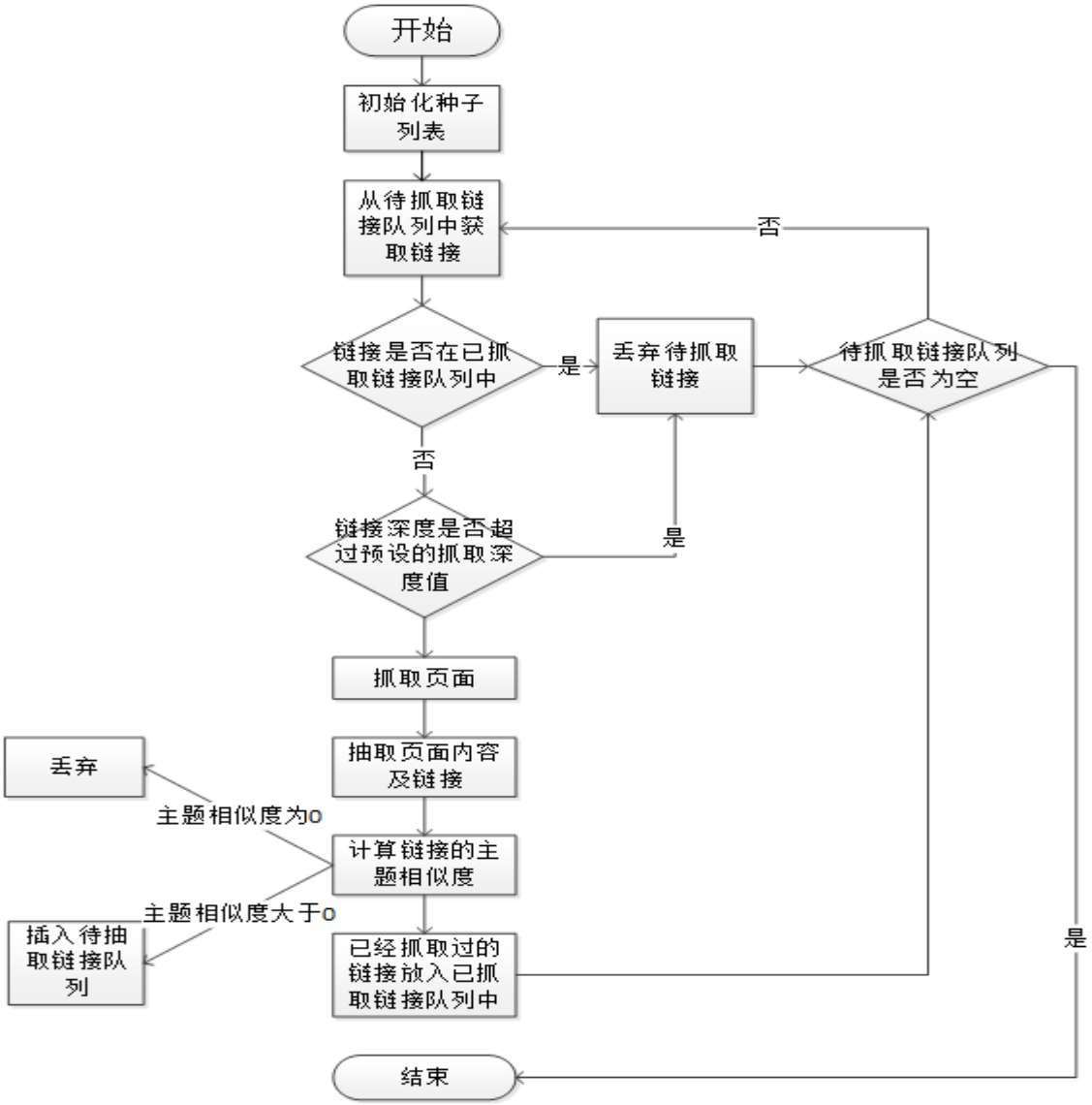

图5-2　基于Best-First算法的搜索策略引导爬虫抓取网页的流程

上图流程具体步骤如下：

(1) 初始化种子列表。此步骤需要将种子列表初始化入待抓取链接列表，此步骤，将种子链接的深度设为 0，主题相似度设为 1。然后还需要预设抓取的深度。

(2) 从待抓取链接队列中获取链接。

(3) 判断待抓取的链接是否在已抓取链接队列。判断待抓取的链接是否在已抓取链接队列中，如果存在则跳到步骤(5)。

(4) 丢弃待抓取链接。

(5) 待抓取链接队列是否为空。判断待抓取列表是否为空，如果为空则结束。

(6) 链接深度是否超过预设的抓取深度值。判断待抓取的链接的深度是否超过预设的抓取深度值，如果超过则跳转到步骤(4)。

(7) 抓取页面。根据链接抓取网页。

(8) 抽取页面内容及链接。

(9) 计算链接的主题相似度。计算抽取出的链接的相似度，丢弃掉主题相似度为0 的链接，将主题相似度大于 0 的链接插入到带抽取链接队列中。

(10) 已经抓取过的链接放入已抓取链接队列中。将抓取过的链接放入已抓取链接队列中，然后跳转到步骤(5)。

5.3.3 实验分析

5.3.3.2 实验环境

本章所使用的实验环境与 3.6.1 小节相同。

5.3.3.2 实验结果与分析

实验的基本思路是：设定步长调节公式(5-3)中的α参数，寻找最优的阈值。

本文选取“汽车”这个主题，选取网易和新浪这两个主流网站的链接作为种子链接。设定搜索深度 depth 为 3，α参数步长为 0.1。

采用收获比的方式评价抓取效率。收获比（Harvest Rate）就是抓取到的网页中主题相关网页数目占所有抓取到的网页总数的比率。相应的计算公式如下：

$$harvestRate = relevant\_pages / pages\_downloadeds \quad (5\text{-}1)$$

其中，pages_downloadeds 代表所抓取到的网页总数，而 relevant_pages 表示这些网页中与主题相关网页的数量。

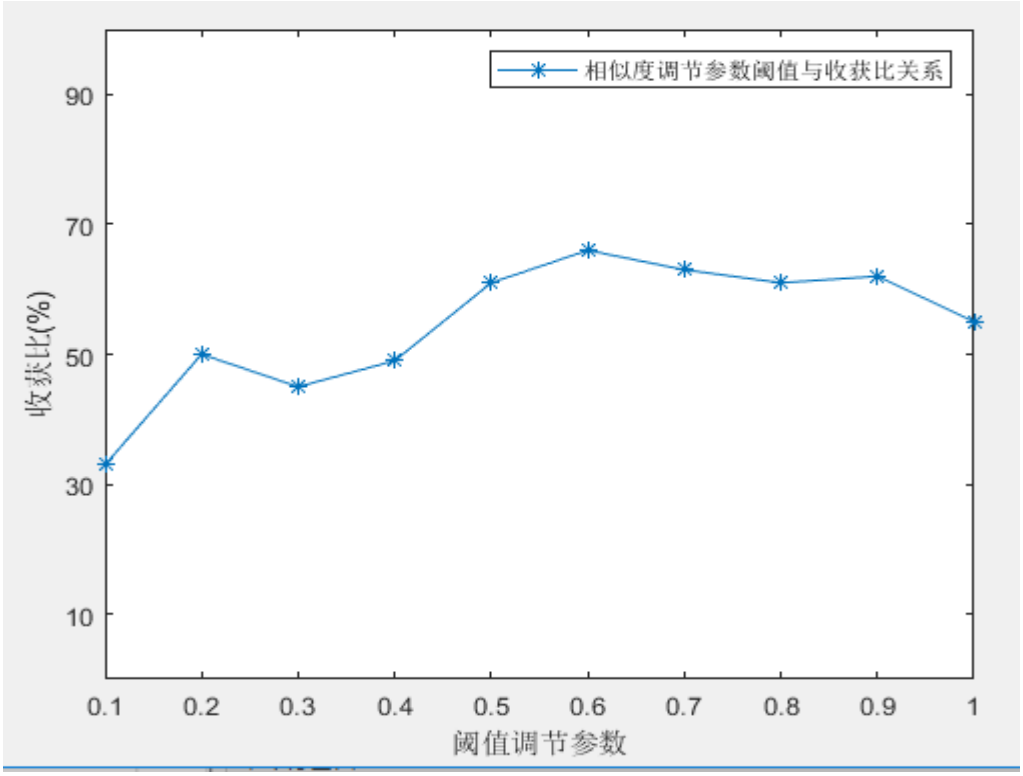

图 5-3　相似度调节参数与收获比关系图

如上图所示，当α取值为 0.6 时能取得较好的效果。

在上述实验的基础上，进一步实验，将传统算法与改进算法进行对比。图 5-4 表示主题爬虫抓取的收获比，横坐标表示抓取的总页面数，纵坐标表示抓取的收获比。

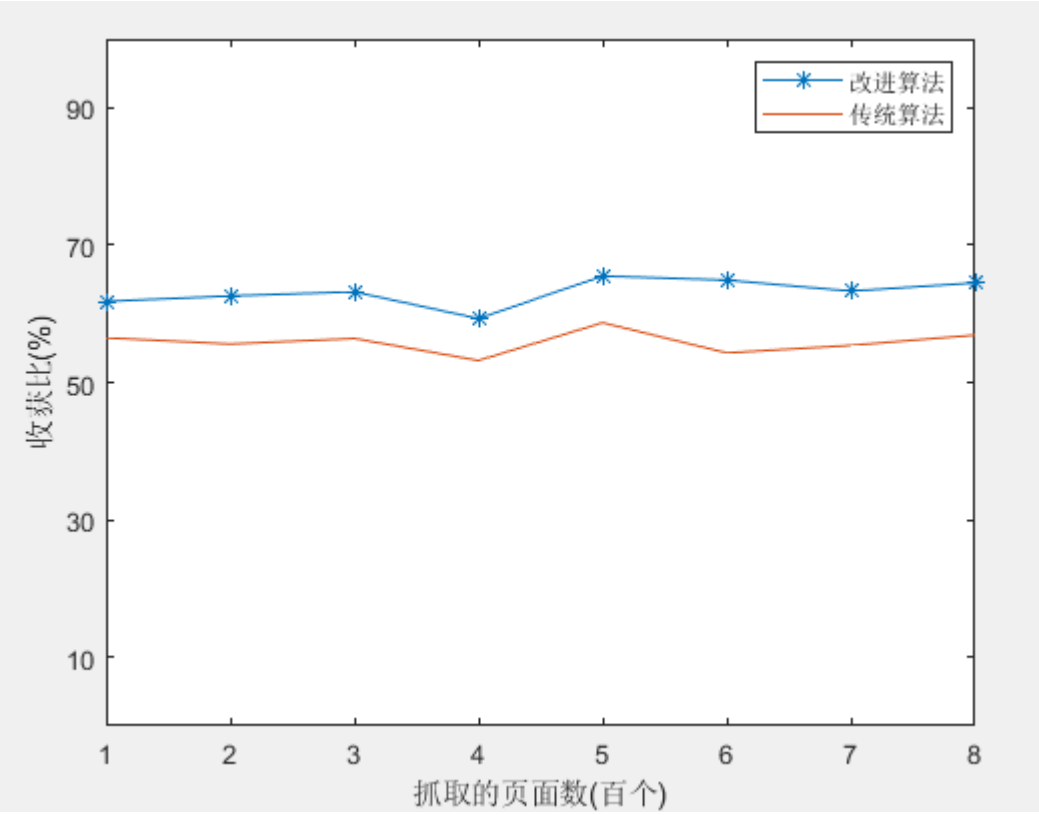

图 5-4　收获比比较

表 5-3 是两种算法的平均收获比比较。可以看到改进算法的平均收获比为 63.1%，比传统算法的收获比的 55.8%高出了 7.3%。

表5-3　平均抓取结果比较

	平均收获比
传统 Best-First 算法	55.8%
改进 Best-First 算法	63.1%

实验结果表明改进的基于 Best-First 算法的主题搜索策略能有效地提高主题爬虫的抓取效率。

5.4　本章小结

本小节主要介绍了主题搜索策略的研究，首先，介绍了通用搜索策略。然后，介绍了目前常用的两种主题搜索策略，并分别介绍了两种搜索策略中比较典型的主题搜索策略的算法。然后详细阐述了本文所用采用的主题搜索策略，并经过实验验证，该主题搜索策略能取得较好的效果。

第六章　总结与展望

6.1　总结

互联网的高速发展使得网络上的信息爆发式增长，信息庞大而杂乱。通用搜索引擎的内容陈旧、查全率查准率偏低、信息冗余等问题越来越不能满足用户特定的需求，所以，针对特定搜索领域的垂直搜索引擎逐渐成为研究热点。而主题网络爬虫是垂直搜索引擎的关键部分，所以对主题网络爬虫进行研究具有较大意义。

本文在现有主题网络爬虫的研究的基础上，进一步进行探索，对其关键技术进行研究。具体的研究内容如下：

(1) 提出了一种网页主题内容抽取的方法。网页主题内容的抽取是网页主题识别的重要步骤，主题内容抽取的准确能提高主题网络爬虫的抓取效率。此方法的基本思路是首先将网页解析成 dom 树结构，然后对网页进行去噪处理，最后根据主题内容在网页中的分布特征抽取出网页中的主题内容。

(2) 本文提出了基于实体链接的主题识别算法。此方法主要将实体链接的技术运用于特征抽取中来提高特征抽取的准确率，进而来提高网页主题识别算法的准确率。首先，利用实体链接的技术抽取出候选特征集合，然后使用信息增益的特征提取方法从候选特征集合中挑选出最终的特征集合，最后使用特征集合来训练朴素贝叶斯分类器并用其对网页主题进行识别。

(3) 最后本文提出了改进的基于 Best-First 算法的主题搜索策略。主题搜索策略的关键点在于链接价值的评估，本文结合网页内容与链接的结构特征来对链接价值进行评估，对基于 Best-First 算法的主题搜索策略进行改进。

6.2　展望

本文所设计的主题网络爬虫在搜索策略及架构方面做了一定的改进，但是还需要进一步改进。

一方面，网页的内容分析上，如何更好的去除网页的噪音，值得进一步研究，如采用“统计学”及“视觉”技术。

另一方面，搜索策略上，可以尝试结合链接结构评价的方式，提高主题网络爬虫的爬行效率。

参考文献

[1] 中国互联网络信息中心.中国互联网络发展状况统计报告(第四次)[R].2017

[2] S.Charkrabani, M.Van den Berg. Focused crawling: a new approach to topic-specific Web resource discovery. In Proceeding of the 8th International World Wide Web Conference, Toronto, CANDA, 1999,545-563.

[3] S.Chakrabani, B.Dom, A.Tomkins. Topic Distillation and Spectral Filtering[J]. Artificial Intelligence Review, 1999, 13(5~6): 409-435.

[4] P.M.E.DeBra, R.D.J.Post.Information Retrieval in the World Wide Web: Making Client-based searching feasible[J].Computer Networks and ISDN Systems,1994,27(2):183-192.

[5] M.Hersovici, A.Heydon, M.Mitzenmacher, etal. The Shark search Algorithm-An application: Tailored Web Site Mapping. Proc of World Wide Conference, Brisbane, 1998[C].

[6] Cho J, Garcia-Molina H, Page L. Efficient crawling through URL ordering[J]. Computer Networks and ISDN Systems, 1998, 30(1): 161-172.

[7] Page L, Brin S , Motwani R. The Page Rank Citation Ranking: Bring Order to the Web[R]. Stanford,CA: Stanford University,1998.

[8] M.Diligenti, F.Goetzee, S.Lawrence. Focused Crawling Using Context Graphs. Proceedings of the 26th International Conference on Very Large Databases, Cairo, 2000[C].

[9] Johnson J, Tsioutsiouliklis K, Giles L, Evolving Strategies for Focused Web Crawling[C]. Proc of Int'I Conf on Machine Learning, 2003, PP.298-305.

[10] J.Rennie, A.K.Mecallum, Using Reinforcement Learning to Spider the Web Efficently[C]. Proceeding of ICML-99, 16th International Conference on Machine Learning, 1999, PP.503-508.

[11] W,Gao, HC.Lee, Y.Miao, Geographically Focused Collaborative Crawling[C], Proceedings of the 15th International Conference on World Wide Web, 2006,PP.287-296.

[12] M.Shokouhi, P.Chubak, Z.Raeesy, Enhancing focused crawling with genetic algorithms[C]. ITCC'05:Proceeding of International Conference on Information Technology: Coding and Computing, IEEE Computer Society, Washington, DC,USA, 2005, PP.503-508.

[13] 陈军，陈竹敏．基于网页分块的 Shark-Search 算法[J]．山东大学学报（理学版），2007,42(9):62-65．

[14] 熊忠阳，史艳，张玉芳．基于信息增益的自适应主题爬行策略[J]．计算机应用研究，2012,29(2):501-504．

[15] 刘强国．主题搜索引擎设计与研究[D]．西安：电子科技大学，2006．

[16] 靳鲁黔，秦颖．独立搜索引擎基本工作原理分析及其简介[J]．农业图书情报刊．2005,17（5）：108-114.

[17] 周鑫．主题 Web 挖掘算法研究与应用[D]．山东师范大学. 2009.

[18] Flake G W, Lawrence S, Giles C L, et al . Self-Organization and Identification of Web Communities[J]. IEEE Computer, 2002, 35(3) : 66-71.

[19] 张航．主题爬虫的实现及其关键技术研究[D]．武汉理工大学,2010．

[20] 高强. 基于 HTML 结构特征的 Web 数据抽取[D]. 南京大学, 2007.

[21] 林子熠，沈备军．基于统计的自动化 Web 新闻正文抽取[J]．计算机应用与软件,2010,27(12)：232-235．

[22] A Lexical Analyzer for HTML and Basic SGML[J].

[23] 罗刚，王振东．自己动手写网络爬虫[M]．北京：清华大学出版社,2010:46-50．

[24] 刘军, 张净. 基于 DOM 的网页主题信息的抽取[J]. 计算机应用与软件, 2010, 27(5):188-190.

[25] 杨文川, 刘健, 于淼. 基于双数组 Trie 树的中文分词词典算法优化研究[J]. 计算机工程与科学，2013, 35(9):127-131.

[26] ICTCLAS2014 汉语分词系统. NLPIR 下载[EB/OL]. 2014.http://ictclas.nlpir.org/.

[27] 陆伟, 武川. 实体链接研究综述[J]. 情报学报,2015(1):105-112.

[28] 舒佳根. 中文实体链接研究[D]. 苏州大学, 2015.

[29] 张涛, 刘康, 赵军. 一种基于图模型的维基概念相似度计算方法及其在实体链接系统中的应用[J]. 中文信息学报, 2015, 29(2):58-67.

[30] 汤效琴, 刘立波, 周涛. 利用海量知识库实现实体标注的一种方法[J]. 计算机工程与科学, 2015, 37(5):895-900.

[31] 谭魏璇. 命名实体与基本名词短语识别研究[D]. 苏州大学, 2010.

[32] 唐博蓉. 基于维基百科的命名实体消歧研究[D]. 北京理工大学, 2011.

[33] Xu B, Xu Y, Liang J, et al. CN-DBpedia: A Never-Ending Chinese Knowledge Extraction System[M]// Advances in Artificial Intelligence: From Theory to Practice. 2017:428-438.

[34] 朱艳辉，徐叶强，王文华等. 中文评论文本观点抽取方法研究[C]. 第三届中文倾向性分析评测论文集. 山东大学：中国科学院计算技术研究所,2011:126-135.

[35] 杨凯峰, 张毅坤, 李燕. 基于文档频率的特征选择方法[J]. 计算机工程, 2010, 36(17):33-35.

[36] 任永功, 杨荣杰, 尹明飞等. 基于信息增益的文本特征选择方法[J]. 计算机科学, 2012,39(11):127-130.

[37] 徐峻岭, 周毓明, 陈林等. 基于互信息的无监督特征选择[J]. 计算机研究与发展, 2012, 49(2):372-382.

[38] 申红, 吕宝粮, 内山将夫等. 文本分类的特征提取方法比较与改进[J]. 计算机仿真, 2006, 23(3):222-224.

[39] 蒋良孝. 朴素贝叶斯分类器及其改进算法研究[D]. 中国地质大学(武汉), 2009.
[40] 杜锐，朱艳辉，鲁琳等. 基于 SVM 的中文微博观点句识别算法[J]. 湖南工业大学学报. 2013(02): 89-93.
[41] 杨仁广，孟祥增．网络多媒体主题搜索策略研究[J]．中国科技资源导刊. 2009,41(2):37-40．
[42] 李学勇．网络蜘蛛搜索策略比较研究[J]．计算机工程,2004,4(4):76-78．
[43] 张丽敏．垂直搜索引擎的主题爬虫策略[J]. 电脑知识与技术.2010,6(15):3962-3963.
[44] 李勇，韩亮．主题搜索引擎中网络爬虫的搜索策略研究[J]．计算机工程与科学,2008,30(3)：4-6．
[45] 陈军，陈竹敏．基于网页分块的 Shark-Search 算法[J]．山东大学学报（理学版）,2007,42(9):62-66．
[46] 喻金平，朱桂祥，梅宏标．基于 Web 链接分析的 HITS 算法研究与改进[J]．计算机工程与应用,2013,49(21):42-45．
[47] 汪涛，樊孝忠．链接分析对主题爬虫的改进[J]．计算机应用,2004,24(s2):174-176．

攻读学位期间主要的研究成果

发表的学术论文：

[1] 马进，朱艳辉，刘璟，田海龙.基于改进朴素贝叶斯算法的主题网页识别的研究[J].信息通信.

[2] 田海龙，朱艳辉，梁韬，马进. 基于三支决策的中文微博观点句识别研究[J]. 山东大学学报（理学版），2014,49(8):58-65. （中文核心，CSCD）

[3] 朱艳辉，田海龙，刘璟，马进. 基于三支决策的新闻情感关键句识别方法[J]. 山西大学学报（自然科学版），2015,38(4):567-572. （中文核心，CSCD）

致　　谢

三年硕士研究生生活即将结束，在这段时间内，我学习到了很多的东西，也认识了对自己很有帮助的老师和同学们。回首研究生生活，自己每天在研究所/食堂和宿舍三点一线式的生活，看起来索然无味，但其中的乐趣只有我自己才能深刻的体会。每一个研究生的背后都有一个为你辛勤付出为你指导的导师，还有一群和你志同道合的师兄师弟，师姐师妹以及玩得好的同学，这些人都值得我感谢。

我最要感谢的就是我的导师朱艳辉教授，导师是一个很慈祥也很负责的人，为我的课题研究给出了很多启示与指导，使我很快进入到研究生的学习和生活中。导师给我的帮助融入到了生活和科研的点点滴滴中，在科研上，朱老师的指导使我进入到了数据挖掘以及算法研究这个全新的领域，自己在这个领域中收获颇丰。在生活中，只要有重大的节日朱老师都会组织聚会、聚餐，使得在外读书的自己，总能找到家的感觉，再次对朱老师说声，由衷的感谢。

感谢周立前院长、满君丰教授为我们提供了良好的学习和科研环境，感谢智能信息处理研究所所长文志强教授、朱文球教授以及智能信息处理研究所的老师们在论文开题、论文中期检查等环节中给予的宝贵意见。

同时要感谢杜锐、鲁琳和梁韬师兄，在我遇到很多不懂的问题时给我的帮助和指导，这些经验使我在研究的道路上少走了很多弯路，感谢刘璟、田海龙，使我有机会跟你们一起学习，同时感谢你们在很多事情上给予我的帮助，和你们在一起学习的日子使我终身难忘。

感谢每一个科研工作者，是你们的成就与成果，让我在这条丰硕的科研之树上继续开花结果。

感谢帮助和关心过我的同学们，和你们一起学习和生活的日子使我永远难忘。

我还要感谢我的父母，是你们给了我大学本科毕业后继续深造的勇气和条件，是父母的辛勤劳动给了我经济上的支撑，每当遇到困难，是父母的安慰和鼓励使我重新继续奋斗。

最后，感谢百忙之中审阅本文而付出辛劳劳动的各位专家、教授！