分类号 学号M20178262

学校代码 10487 密级



**硕士学位论文**

**面向招聘领域的垂直搜索引擎系统的设计与实现**

|  |  |
| --- | --- |
| 学位申请人： | 方 聪 |
| 学科专业： | 软件工程 |
| 指导教师： | 卢力 副教授 |
| 答辩日期： | 2019.12.18 |

**A Dissertation Submitted in Partial Fulfillment of the Requirements**

**for the Degree of Master of Engineering**

**Research on optimization method of click rate estimation in recommendation system**

|  |  |
| --- | --- |
| **Candidate :** | **Fang Cong** |
| **Major :** | **Software Engeering** |
| **Supervisor :** | **Prof. Lu Li** |
|  |  |

**Huazhong University of Science & Technology**

**Wuhan 430074, P. R. China**

**May, 2019**

独创性声明

本人声明所呈交的学位论文是我个人在导师指导下进行的研究工作及取得的研究成果。尽我所知，除文中已经标明引用的内容外，本论文不包含任何其他个人或集体已经发表或撰写过的研究成果。对本文的研究做出贡献的个人和集体，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。

学位论文作者签名：

日期： 2019年 12月 18日

学位论文版权使用授权书

本学位论文作者完全了解学校有关保留、使用学位论文的规定，即：学校有权保留并向国家有关部门或机构送交论文的复印件和电子版，允许论文被查阅和借阅。本人授权华中科技大学可以将本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和汇编本学位论文。

保密□， 在 年解密后适用本授权书。

本论文属于

不保密□√。

（请在以上方框内打“√”）

学位论文作者签名： 指导教师签名：

日期： 年 月 日 日期： 年 月 日

# 摘 要

随着互联网的飞速发展以及大数据的到来，各个领域充斥着大量的信息。如何检索出有用的信息成为了各领域的首要任务。因此，结合垂直搜索引擎技术开发出招聘信息的搜索引擎平台成为了该领域的亟需解决的问题。

主题网络爬虫策略融合了网页采集、网页分块、主题相关度判定以及信息提取等技术。该策略在领域网页采集方面有着明显的优势，如具有很高的网络资源利用率以及信息采集的准确率。本文在考虑网页的相关度计算中，融合了链接所在的内容块与主题的相似度，通过对链接的相似度与网页内容的相似度的权重组合，来过滤与主题无关的网页，同时链接与主题相似度也能够降低爬虫服务对无关链接的访问量。

本文结合全文检索框架 Lucene、分布式爬虫框架以及 HBase 集群开发了一套招聘信息的垂直搜索引擎。该爬虫框架实时抓取国内多个招聘网站数据，目前已经抓取的网页库达到千万级的招聘信息数据，并存储到 HBase 集群，可用于数据分析以及挖掘等。该搜索引擎融合了同义词以及 Rocchio 相关反馈算法对原始查询的扩展，进而优化 Lucene 的默认查询，并提供了域查询以及模糊查询等多样化的功能。

最后对招聘领域的垂直搜索引擎中的网络爬虫以及检索模块进行了功能测试，并与通用搜索引擎的检索效果进行对比实验，本系统的检索结果具有很高的准确率。同时也对扩展查询与域查询进行了测试，结果表明扩展查询具有更高的召回率。

**关键词： 垂直 搜索引擎；主题网络爬虫；Lucene；相关反馈算法**

# ABSTRACT

The click rate prediction has been widely used in the current computing advertising and recommendation systems, and the advertising business is also the core realizing business of most companies. Therefore, the CTR estimation has received more and more enterprise attention and application research.

In recent years, deep learning has developed rapidly, and the application of the industry has become more and more mature. A great advantage of deep learning is that it can deal with nonlinear related problems well, and can mine the feature combination sequence that ordinary machine learning algorithms cannot mine. And excellent in the results of real data testing.

The industry is currently applying a deep learning method to CTR estimates. There are also many achievements, such as Wide&Deep, DNN, and FM, but limited to the diversity of current deep learning frameworks. There is no unified theory, but the goal is to use new frameworks. , through experimental methods to detect whether it can improve the accuracy of CTR estimation.

Therefore, in combination with my internal CTR estimation model during the internship period of the company, I hope to be able to achieve CTR estimation in the optimization method of the recommendation system, and work hard to solve the problem that can be solved at present. A new approach to the problem, to reduce the company's investment in the field and improve the accuracy of CTR estimates, and to keep up with the rapid development of the deep learning framework, improve the accuracy of CTR sequencing in the department's internal recommendation system.

**Key words： Click rate prediction, deep learning, optimization method research, recommendation system**

目 录

[摘 要 4](#_Toc23271985)

[ABSTRACT 5](#_Toc23271986)

[参考文献 11](#_Toc23271987)

1. **绪论**
   1. 研究的背景及意义

随着互联网的飞速发展，人们的生活中出现了越来越多的智能化产品，而这些产品产生了不同类型的信息，同时也在不同程度上改变人们的生活方式。根据中国互联网络信息中心最新报告表明[1]：截至2018年12月，我国网民规模达8.29亿，普及率达59.6%，较2017年底提升3.8个百分点，全年新增网民5653万。我国手机网民规模达8.17亿，网民通过手机接入互联网的比例高达98.6%。

面对日益增长的信息，如何能快速的检索到所关注的信息成为了关键。搜索引擎的出现在很大程度上解决了该问题，它为互联网用户提供了一个快速检索信息的平台[3]。搜索引擎旨在利用从互联网上采集到信息资源，进而提供信息查询的服务。目前市场上出现了很多搜索引擎，比如谷歌、百度，涵盖了多样化的信息，并能够根据用户意图检索出相关的信息。目前，搜索引擎成为了人们用于信息检索的最为常用的工具。

当前“互联网+”时代下，人才招聘的渠道越来越广泛。 在我国改革开放初期，企业招聘大多采用报纸招聘、人才市场定点招聘和电视广告招聘形式。随着科技的发展，陆续出现了许多大型的招聘平台，企业和人才通过平台实现信息的交互，扩大了企业招聘渠道。近年来，“互联网+”时代的到来，招聘渠道实现了进一步扩展，通过网络能够实现人才与企业之间的面对面交流；通过互联网与手机终端的结合，实现企业信息的广泛发布。

通用搜索引擎可以为互联网用户提供查询服务，例如百度、谷歌[4]、雅虎等，但近几年通用搜索引擎面临着来自领域大数据带来了巨大的挑战[5,6]，因为通用搜索引擎旨在为所有用户提供服务，而面对特定领域的信息检索，可能造成搜索结果不匹配，因此通用搜索引擎不能满足用户对特定领域信息的搜索需求[7,8]。针对这类问题，各个网站利用自己的信息源来提供信息的检索功能，在某种方面上改善了这类问题。例如在企业招聘领域当中，各大招聘网站都有提供招聘信息的检索功能，但这些招聘网站的信息并不做到数据共享，使得应聘者仍然需要在各大招聘网站反复检索，才能检索到满足自己需求的招聘信息。所以，对于校招以及社招的应聘者来说，一个亟待解决的问题也随之而来，即如何在这海量的招聘信息中快速并准确的定位到自己所需要的招聘信息。

垂直搜索引擎能够针对性的处理以上问题。因为垂直搜索引擎可以针对特定的领域或者该领域中某一模块进行专业和深入的分析、挖掘以及定位，更为精准的提供有一定价值的信息和相关服务，有效地弥补了通用搜索引擎对特定领域中信息检索不准确的缺陷。因而，如何利用垂直搜索引擎的相关技术，为招聘领域构建一套招聘信息的垂直搜索引擎成为本文所要解决的问题。

* 1. 国内外研究现状

在这个大数据的时代，对于垂直搜索引擎等相关技术的研究仍然是社会各界的热点方向，越来越多研究者们结合人工智能的知识，进而对垂直搜索引擎进行改进。大数据的到来同样使得人们对快速准确地检索所需信息的要求越来越高，因此各种搜索引擎不断更新换代。正是由于这两个方面的目标与需求，社会各界人士都参与了垂直搜索引擎的研究中。如 Yue[9]针对数据急速增长的情况，提出了金融领域的垂直搜索引擎的设计和实现，该搜索引擎能很好的满足金融领域的实际需求。Xu[10]等人设计了基于垂直搜索引擎的论文提交推荐系统，通过改进基于网页链接的过滤算法和改进基于内容的过滤算法提高了搜索的准确性。WangWei[11]旨在对海量的网络数据进行提取和分类，建立了有效的垂直搜索引擎。Cheng[12]等人设计和开发了一个基于 Hadoop 的垂直搜索引擎 HVSE，虽然该引擎是基于传统搜索引擎的基本原理，但其对目前面向主题的爬虫算法进行了改进，并结合 MapReduce 编程模型进行数据处理。Konstantin[13]描述了一种自动集成Web 查询接口的方法。该方法允许用户通过统一的 Web 查询接口以透明方式查询特定域中的多个网页数据库。Sejal[14]通过将垂直搜索引擎应用到图像推荐领域，基于余弦相似图像推荐框架以及文本搜索，对原始查询进行扩展，进而返回结果集。Hong[15]通过对专业搜索引擎的研究，并结合改进的网页排名算法，进而

设计并实现了一个专业的垂直搜索引擎。Nikulin[16]公开了一种生成搜索引擎结果页面的方法，融合垂直搜索引擎以及历史交互数据，使得搜索结果页面的排序效果更好。English[17]有效的设计了一种旅行预订搜索引擎，该系统根据一个或多个约束构造双重查询，搜索出双重结果集，并且第一个结果集具有很高的购买率。

同样在工业界也相继出现了很多垂直搜索引擎的产品，例如常见的国外垂直搜索引擎有航班搜索引擎 SkyScanner、博客搜索引擎 Ice Rocket、图片搜索引擎TinEye、音乐搜索引擎 MixTurtle、域名搜索引擎 Panabee 等。其中，SkyScanner是面向航班领域的垂直搜索引擎，旨在提供对航班的信息的检索服务[18]，IceRocket 旨在针对 Twitter、博客、Facebook 等站点进行专业化的搜索[19]，MixTurtle 作为国外最好的音乐搜索引擎。同样科研机构也开发了一系列的学术领域的垂直搜索引擎系统，例如，citesee 是计算机科学领域的一个非常著名的检索系统，由NEC 研究院研发而成，其核心是自动索引与分类电子文件[20]。IRISWeb系统是由计算机科学系和北卡罗莱纳大学法学院合作开发，可以检索互联网上使用自然语言全文法律信息，使检索效率大大提高[21]。国家科学数字图书馆是一套集科学、数学、工程和技术的大型在线数字图书检索系统[22,23]。

国内早在 2005 年也加入了垂直搜索引擎领域的研究中，比较有名的垂直搜索引擎网站有百度学术、亨者搜索、美团、去哪儿、奇虎、搜狗、职友集、咕嘟妈咪等。其中，百度学术是一个专业的面向学术搜索引擎，美团是一个集娱乐、美食于一体的搜索引擎系统，去哪儿是目前较为流行的面向旅游搜索引擎之一，而房天下是一个房产领域的垂直搜索引擎。同样在招聘领域也有相应的垂直搜索引擎产品，如智联招聘、前程无忧、猎聘网、BOOS直聘、拉勾网、58同城等专业招聘搜索引擎。

如今，大量的招聘搜索引擎提供相同或相似的功能，即一个房地产搜索引擎根据用户指定的一个地点，行业，职能，薪资等，搜索出相似的招聘信息。类似前程无忧、拉勾网、智联招聘等招聘网站搜索平台，由于单个招聘平台的信息不足以及招聘平台的多样性，从而应聘者需要到各个招聘网站平台上搜索满足自己需求的职位，这将直接导致应聘者时间的浪费，同时也影响了应聘者的体验。所以对于应聘者来说，对现有二手房资源进行数据整合成为了关键。为有效利用已有的招聘信息资源来发挥其价值，能否利用垂直搜索引擎相关技术，如网页爬虫、网页分类、分词算法等，设计一个高覆盖率的招聘信息垂直搜索引擎平台，从而使得招聘垂直搜索引擎更为智能化与人性化，也方便应聘者找到自己满意的工作。本文针对以上搜索引擎和招聘信息资源存在的问题，展开了招聘领域的垂直搜索引擎平台的设计与实现。

* 1. 主要内容及结构安排

论文首先概述了垂直搜索引擎的研究背景和意义、垂直搜索引擎以及招聘领域的特点以及相关技术。其次，设计了网络爬虫模块中的网页抓取策略，利用网页分块技术提取链接所在的内容块，并结合链接及网页与主题相似度来过滤无关链接以及页面，进而实现招聘主题爬虫系统。然后运用分布式系统基础架构Hadoop 以及分布式数据库 HBase 建立分布式存储平台，结合 HBase 分布式数据库的特性，设计表结构，从而实现海量数据存储。最后结合全文检索框架 Lucene以及中文分词器 IKAnalyzer 建立二手房索引库，同时融合同义词以及 Rocchio 算法对全文检索框架 Lucene 进行查询优化，并实现了一个二手房的垂直搜索引擎。

本文有五个组成部分，各个部分如下：

第 1 章绪论，对招聘信息的垂直搜索引擎的背景及意义进行详细的阐述，接着对国内外在垂直搜索引擎领域的研究与工作情况进行阐述与分析，最后对整篇文章的结构与编排进行描述。

第 2 章垂直搜索引擎相关技术，首先对搜索引擎的工作原理进行了介绍，同时结合招聘领域的专业知识，分析二手房的垂直搜索引擎的特性。接着对网络爬虫技术中的主题爬虫框架与爬虫策略进行介绍。再对目前中文分词的主流算法进行介绍，同时详细阐述了非关系型数据库 HBase 涉及到的相关知识，如基本架构、特性等。最后，对全文检索框架 Lucene 的架构、索引等组件进行了描述。

第 3 章招聘的主题网络爬虫策略设计，主要介绍主题网络爬虫策略所涉及到的技术与算法，如链接遍历搜索策略、内容块相似度计算以及基于页面内容的网络爬虫策略。最后本文采用融合链接与页面内容的网络爬虫策略来实现二手房的主题网络爬虫服务，通过计算链接所在内容块与主题的相似度以及网页与主题相似度来保证招聘的主题网络爬虫服务抓取的网页具有很高的主题覆盖率。

第 4 章基于 Lucene 检索模型的查询优化，主要对基于 Lucene 检索模型的检索问题、Lucene 检索模型、Rocchio 相关反馈算法做了详细的介绍。最后，本文通过结合同义词以及 Rocchio 算法，在 Lucene 检索模型的基础上扩展查询向量进而优化查询功能。

第 5 章招聘的垂直搜索引擎的设计与实现，首先对招聘信息的垂直搜索引擎的功能、目标以及整体设计做了详细描述。其次，根据该系统的功能模块进行详细设计，主要包括网页采集模块、索引构建模块以及检索模块。最后，对各个模块进行实现，并给出各个模块的实现效果图。

第 6 章系统测试，主要是对网络爬虫以及检索模型进行功能测试。首先对国内多个招聘网站中的招聘模块进行爬虫测试，并生成测试报告。最后，对检索模块中的默认查询、扩展查询以及域查询等功能进行测试。

1. **垂直搜索引擎相关技术**

**2.1搜索引擎的工作原理**

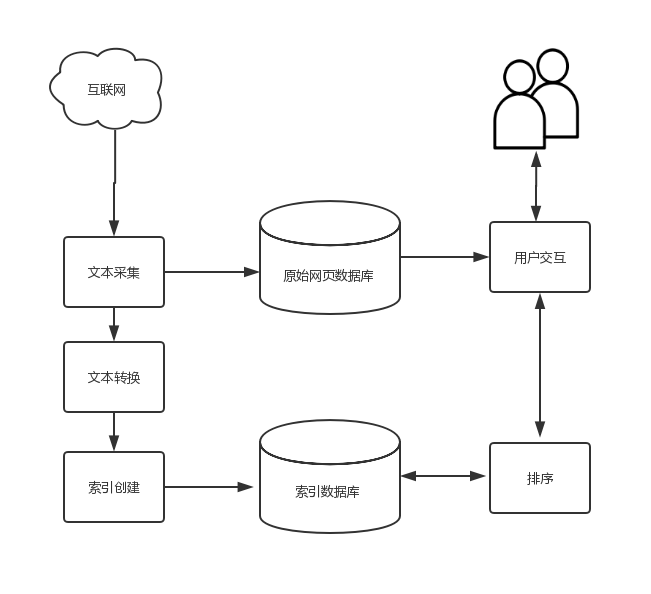
搜索引擎旨在通过搜集互联网上网页信息，进而给用户提供信息检索的平台[24]。由此可见，搜索引擎主要包含索引处理和查询处理两个组件，其中，索引处理组件的目的就是搜集互联网上的网页信息，并存储到索引库中。当然在这个过程中包含着许多文档处理、转换等操作。而查询处理组件旨在通过索引组件中的索引库，对用户提供的查询词进行查询操作，并将查询结果集采用特定排序算法返回给用户，同样的是查询处理的过程中也包含着查询词处理、日志采集等操作，其中日志数据可以用于信息反馈，进而使得检索结果能够反映出了用户的喜好等信息[25]。上述只是简单的规划为两个组件，其实可以更为细致的划分，通用的搜索引擎主要有文本采集模块、文本转换模块、索引模块、用户交互、排序模块，其基本结构如图 2-1 所示。

图2-1搜索引擎基本结构图

（1）文本采集模块

文本采集模块主要功能是通过网络爬虫服务对互联网上的网页信息进行抓取并存储到原始网页数据库中。而在文本采集模块又涉及到文档信息源的设定、网页链接遍历策略设定、网络爬虫架构、文档存储方式以及更新策略等一系列的子模块。其中，文档信息源的选定决定该模块抓取怎样的数据源，网页链接遍历策略主要指的是网络爬虫链接访问规则，文档存储介质可以分成关系型数据库和非关系型数据库，最后更新策略决定网页信息的时效性[26]，这些因素都会影响文本采集模块的性能。

（2）文本转换模块

文本转换模块指的是将原始网页信息转换成特征词以及链接等有用的信息。而在文本转换模块中包含网页解析、停用词去除、词干提取、超链接的提取与分析、信息抽取、分类器等部分，原始网页信息经过文本转换模块中的一系列的处理，得到原始网页的词信息、链接信息以及类别信息等有价值的信息，以供后续索引模块使用。

（3）索引模块

索引模块旨在对预处理后的文档进行汇总，并对记录词、特征以及文档进行统计，结合文档词项信息构建倒排索引[27]。而倒排索引的结构方式会影响检索模块的检索方式，倒排索引为了降低检索模型中的复杂度，会对每个索引项的权重信息以及文档的权重信息进行存储。如果需要将索引分发给多台服务器，则还需要索引分发组件，通过网络中每一个节点都维护着自己的索引表以及文档集合，这样可以降低网络通信延迟，进而提高效率。

（4）用户交互

用户交互旨在处理用户提交过来的查询请求，它主要负责对原始查询词进行拼写检查以及查询扩展等转换操作，然后提交给检索组件并将查询结果集展示给用户；同时用户也可以将搜索出来的结果是否满足要求反馈给搜索引擎[28]。

（5）排序模块

排序模块旨在对查询结果按照特定的评分机制进行评分排序[29]，进而使其排好序的结果集更有利于用户浏览，以及快速定位到用户所需的信息，而不是需要转到后面几页才能找到用户想查找的信息。

**2.2 招聘信息的垂直搜索引擎的特点**

对于构建二手房的垂直搜索引擎平台涉及到对垂直搜索引擎以及二手房领域信息进行分析并提取其特点，才能够使其抓住这两者的共性，进而在两方面上满足需求。下面对各个方面进行描述及分析。

**2.2.1 垂直搜索引擎的特点**

相比于通用搜索引擎的特点，垂直搜索引擎侧重于某一领域的信息检索，它的数据量以及信息种类多样性都相差甚多，正是由于这个原因，垂直搜索引擎更为侧重的是信息的准确性以及完整性。如果垂直搜索引擎检索出来的大部分信息都是与领域无关的，那么会降低整个垂直搜索引擎的用户体验。根据领域相关性的特性，总结出垂直搜索引擎的特点如下所示。

（1）面向领域的网页信息

对于垂直搜索引擎而言，只需要抓取特定领域的网页数据，所以在网页的抓取策略中，需要融合对链接以及网页的相关性计算，进而过滤无关的网页。抓取的网页与主题相关程度越高，则检索效果也就越高。除了加入链接与网页的主题相关性的判断之外，设置与主题相关的初始地址同样也助于主题爬虫的准确率，而这些初始地址可以通过人工选择来实现。

（2）领域化的中文分词

由于垂直搜索引擎的目的就是对某个领域的信息进行采集、处理、检索等操作，而这些操作都涉及到中文分词[31]，其中，中文分词技术仍然还是垂直搜索引擎技术中的一大难题。如何对专业词汇的切分处理成为中文分词的关键。因此垂直搜索引擎涉及到的中文分词模块需要该领域的专业词汇，并不是通用的词库。所以构建出完善的领域词库有利于改善垂直搜索引擎的中文分词效果。

（3）结构化信息的提取

由于垂直搜索引擎能够提供个性化的查询功能，如域检索、全文检索、域排序等功能，所以垂直搜索引擎需要对各个关键域信息的解析，如招聘领域的“行业”、“职能”、“薪资”、“地点”等信息。

（4）个性化的复合检索以及排序方式

垂直搜索引擎在检索功能上需要提供丰富的个性化功能，同时在排序方面也同样需要更为多样化的方式，不再仅仅依靠文档的评分，可以依靠域属性进行排序。这些个性化功能使垂直搜索引擎更能满足用户的查询需求。

**2.2.2 招聘信息的特点**

根据对招聘领域的相关专业知识以及互联网上的招聘信息的分析，其主要包含三个方面：其一是招聘领域数据本身含有较多的专业词汇，而这些专业词汇分词的好坏直接影响网页主题的判定。其二，由于目前招聘网站平台众多，企业很大可能在多个平台发布同一个招聘信息，而这会产生大量的重复的招聘领域的网页信息。其三，招聘信息的实时性要求较高，过期的招聘信息对于应聘者来说没有意义，因此要过滤掉过期的招聘信息。

**2.3网络爬虫技术**

**2.3.1主题爬虫架构**

主题爬虫架构融入了在网页抓取过程中过滤与主题无关网页的判定流程[32]，其他的流程与通用的爬虫框架构大体一致。对于通用的爬虫架构而言，它将存储所有的原始网页信息。主题爬虫中的过滤流程主要分为两部分，其一是通过事先设定好的 URL 访问规则，将符合特定的 URL 访问规则的 URL 存放到待爬取URL 队列中，反之，则直接过滤掉。其二是将原始网页解析出特定区域的页面内容，并结合主题相关性分析计算出该网页与主题的相关度，同时与某一阈值进行比较，确定是否存储该网页。主题爬虫架构如图 2-2 所示。

//图片

2.3.2 网络爬虫策略

垂直搜索引擎的网络爬虫策略旨在抓取具有较高的主题相关度的网页，所以它可以加入不同的限定规则来优化网络爬虫策略[33]。下面介绍常用的策略主要有以下两种策略。

（1）域名预选策略

域名预选策略，是通过限制与主题相关的域名来抓取网页信息，这样的策略可以有效的保证网页与主题的相关度，但是它的缺点就是抓取的网页信息不够全面。由于域名预选策略所抓取的网页有着网页结构单一、明确等特点，所以有利于后续的网页解析。

（2）基于相关度的策略

基于相关度的策略，是在抓取网页的过程中，通过计算链接的相似度进而确定该链接是否存储到待访问的地址队列中，同时也会计算网页的相似度来选择是否保存网页信息。这种策略不仅能加快网络爬虫速度以及提升抓取网页的纯度，而且能在一定程度上降低无关网页的访问量以及存储量[34]。由于加入网页相关度的计算，所以相比于域名预选策略而言，复杂度较高，并且一定时间内抓取的网页数量会减少很多。

2.4 中文分词技术

2.4.1 中文分词概述

中文分词作为自然语言处理的一个基础模块，分好的词集可以应用到自然语言处理、机器学习等多个领域中。相比于英文以及其他语言的分词难度而言，中文分词显得更为棘手。原因在于英文中每个词之间是通过空格来区别开的，而在中文词语之间是紧密相连的，除了标点符号之外；同时在中文里大多数词语在不同语境下可能表达着不同的意思。正是出于这样的原因，目前大多数应用都是根据不同需求制定不同语法规则以及中文词库等限制来保证满意的分词效果。 目前，由于机器学习、深度学习的技术日渐成熟，开始引入了人工智能算法来提高中文分词的准确率。如隐马尔科夫模型、条件随机场、支持向量机、深度学习等算法。不少的研究机构以及高校都在基于机器学习算法来实现中文分词。例如，哈工大的“语言云”，以语音技术平台为基础，提供中文分词、词性标注等接口，斯坦福的基于条件随机场算法的分词器等。下面对中文分词算法的种类进行详细介绍。

2.4.2 中文分词基本算法

现有的中文分词算法都是基于不同分词规则以及语料库。不同的分词算法具有不同的应用场景。接下来依次详细介绍几种基本算法。

2.4.2.1 基于词典的中文分词算法

基于词典的匹配算法仍然是目前主流的中文分词方法。词典是汉语自动分词的基础，它的结构直接关系到分词的速度和效率[35]。自动分词是中文信息处理系统的基础，根据分词结果进而对汉语文本进行句法分析和语义分析。词典会直接影响切分速度。字典的数据结构主要是通过索引方法，包括索引表、倒排表、哈希表和搜索树[36]。该算法的思想是将提取出的连续字符与词典进行匹配，如果能够匹配到词语，则将其作为一个词语切分出来，而提取连续字符的规则存在多种方式。下面介绍几种基于词典的分词算法[37]。

（1）正向最大匹配算法

正向最大匹配算法，简称 MM 分词法。其主要思想就是从左到右将待分词文本中的 MaxLen 个连续字符与词表匹配，其中 MaxLen 为词典中汉字最多的词条长度，如果待分字符与词表匹配上，则将该连续字符切分出一个词语。如果没有匹配上，则删掉该连续字符的最后一个字符，进而生成一个新的词重新再与词典中的词条进行匹配，依次重复以上操作，直到剩下最后一个字符，将其切分出一个单字。接着剩余的字符按照上述操作依次执行下去，直到整个待分词文本切分完成[38]。正向最大匹配法流程图如图 2-3 所示。

图片

（2）逆向最大匹配算法

逆向最大匹配算法，简称 RMM 分词法。其分词原理是每次从待分词文本的尾部开始选取 MaxLen 长度的连续字符与词典进行匹配，如果匹配上，则将该连续字符切分成词语，如果匹配失败，则去掉该连续字符的首个字符重新与词典匹配，直到待分词文本切分完为止[39]。其思想与正向最大匹配算法正好相反。

（3）最佳匹配分词法

最佳匹配分词法，简称 OM 分词法，就是结合正向最大匹配法和逆向最大匹配法的分词效果进行比较，进而确定分词结果[40]。如果两者分词结果重合时，则该分词结果表示为正确的，如果是两者分词结果不一致时，则根据一些启发式规则确定分词结果，如取分词数量较少的结果作为分词结果等。 根据统计信息可得，逆向匹配的切分精度要好于正向匹配，处理歧义的地方不太多。由于基于词典的中文分词算法具有复杂度较低、实现简单等特性，所以它被广泛应用在实际项目中。当然需要结合领域的语言信息，进而将其精度提高到项目所需的要求。

2.4.2.2 基于理解的中文分词算法

基于理解的分词算法旨在利用语法分析以及语义分析来实现对词的歧义的判定，从而获得更好的分词效果[43]。但该分词算法需要大量而又复杂的语法知识与规则，所以目前该分词算法并没有取得很好的准确率。该算法主要分为专家系统分词法以及神经网络分词法[44]。同时为了取得更好的分词效果，可以通过整合这两者来提升准确率。

2.5 非关系型数据库 HBase

2.5.1 HBase 概述

非关系型数据库 HBase，是在分布式文件系统 Hadoop[45]基础上实现的一款非关系型数据库，并且它摒弃了关系型数据库的事务等特性，正是由于这种设计方式，它具有高可靠性、面向列、高性能、可伸缩等特性[46]。在面向处理海量数据业务时，HBase 集群可以通过 MapReduce 来实现并行化，进而提高数据处理速度。不同于传统的关系型数据库的表结构，HBase 数据库中的表结构是由行键、列族、列名以及时间戳构成，而对于每个存储数据由这四种类型的信息组成[47]。同时 HBase 数据库并没有限制存储数据的种类，允许动态存储不同类型的数据。由于 HBase 数据库是分布式的，则它需要 ZooKeeper 来调节各个节点服务器的工作情况。除了数据表结构中的扩展，HBase 数据库集群同时也可以支持节点服务器以及数据存储的横向扩展。

2.5.2 HBase 特性

相比于传统的关系型数据库，在处理海量数据的业务时，HBase 在很多方面都表现的很好，如海量存储、并高发等，而这些优点归功于它的几个特性，如图

2-4 所示。

图片

（1）海量存储

HBase 能够做到对 PB 量级的数据进行存储，而且对于 PB 量级的数据进行查询，能在几十到几百毫秒内返回该数据信息[48]。而拥有这么好的性能归功于HBase 的极易扩展性。仅仅通过添加从节点服务器就可以扩展 HBase 的存储性能，同时也可以提升数据处理性能。

（2）列式存储

HBase 采用的是列族存储，HBase 表中包含多个列族，而列族又包含多个列，

每个列存储着详细数据，在创建表的时候就必须将列族确定好。为了与关系型数

据库的表结构进行比较，看出 HBase 的表结构的不同之处，下面对一个关系型数

据库的表和 HBase 数据库的表进行描述，如表 2-1、2-2 所示。

表2.1

表2.2

（3）极易扩展

HBase 的扩展性分为两个方面，一个是基于节点服务器上的扩展，一个是基于存储的扩展。通过横向添加节点服务器进行水平扩展，提升集群的数据处理能力，同时提升集群处理分区的能力。节点服务器的作用就是管理分区、承接业务的访问，这个后面会做详细的介绍。其次通过横向添加数据节点，进而对存储方面进行扩容，提升集群中数据存储能力以及后端存储的读写能力[49]。

（4）高并发

由于目前大部分使用 HBase 集群架构，都是采用廉价服务器来搭建的，因此单个节点服务器读写的延迟其实并不小，一般在几十到上百毫秒之间。但是在并发非常高的时候，HBase的单个节点服务器上读写延迟却下降的很少。因此HBase集群能维持一个低延迟、高并发的状态。

（5）稀疏

稀疏主要是针对 HBase 列的灵活性，在列族中，你可以指定任意多的列，在列数据为空的情况下，是不会占用存储空间的。

2.5.3 HBase 集群架构

HBase 集群采用主从服务器架构搭建的，并且需要由分布式文件系统 Hadoop集群来支撑，同时该集群还需要 ZooKeeper 来负责各个服务器相互协同工作[50]。总体结构如图 2-5 所示。

图片

从上面的架构图可以清晰地看出 HBase 中的各个组件的相互关系，下面详细介绍各个组件的功能，如下所示。

其中主节点的详细功能如下所示。

1）管理从节点服务器，实现其负载均衡。

2）管理和分配分区，主要是处理在分区合并、分裂以及分区所在的服务器宕机时，完成对分区数据的转移。

3）实现数据库定义语言操作。

4）管理命名空间和表的元数据，实现在数据查询过程中，快速定义到数据所在的从节点服务器以及分区。

5）权限控制。

从节点的详细功能如下所示。

1）对本地分区的读写与管理。

2）读写分布式存储文件并管理表中的数据。

3）接受从客户端提交过来的查询请求，直接通过从节点中读写数据。 ZooKeeper 集群主要是用作调节各个节点服务器的任务分配、状态信息等。功能如下所示。

1）维护着元数据信息以及协调元数据的更新。

2）实现在主服务器宕机时从服务器代替为主服务器提供服务，进而做到容灾机制。

2.6 全文检索框架 Lucene

2.6.1 Lucene 概述

Lucene 作为一个开源全文检索框架，主要应用于构建全文检索引擎，它仅仅是提供给开发者用来搭建搜索引擎的框架[51]。全文检索框架 Lucene 由查询组件和索引组件两部分组成，其中，这两部分都已提供了完善的调用接口。全文检索是用于在全文数据库中搜索单个计算机存储的文档或集合的技术。一个完备的全文检索系统必不可少的两个基础功能是索引建立和索引搜索，此外还需要考虑到系统的二次开发接口、用户调用接口等。全文检索系统的主要功能基本上可以概括为建立索引，根据索引查询得到结果，对原始索引进行增加或者优化结构等功能。从结构上考虑，全文检索系统的核心主要由四部分构成，分别是用于构建索引的索引引擎、用于查询的查询引擎、用于分析文本的文本分析引擎以及针对外部功能模块的各种接口，这些与第三方的软件或者系统共同组成一个完整的全文检索系统。全文检索系统的系统结构图如图 2-6 所示[52]。

图片

2.6.2 Lucene 系统结构

Apache Lucene 是一个基于 Java 的开源的检索框架，已提供通用搜索以及与搜索相关的功能，如索引、高亮、词语分析等[53]，并提供给应用程序调用的 Java接口。Apache Lucene 主要有数据存储、索引、查询、分词器、文档等几个部分，基于面向对象思想对这几个对象进行设计，确定彼此之间的继承关系。Apache Lucene 不仅设计该全文检索模型，并对其进行了实现，如基于 Java、Net 等。Lucene 的系统结构涉及到一系列的组件，这些组件都由不同实现类来实现的。具体系统结构与源码组织详情如图 2-7 所示。

图片

如图 2-7 所示，全文检索框架 Lucene 的系统结构分为索引文件、查询处理组件、分词器以及基础结构封装等几个部分，而且每个部分都有不同的类来实现[30]。其中，索引文件是由 store 对象以及 index 对象来实现的，查询处理组件是由 search 对象来实现的，而对于查询词的分词操作是由 analysis 对象来实现的，文档数据的封装由 document 对象来实现。不仅仅是简单的实现，每个对象互相之间还有调用关系，进而完成整个检索流程。 对于用户输入的查询词，需要查询解析器来将查询词进行分词、语法分析以及查询扩展等操作，这些有的可以通过 queryparser 对象来处理，有的则需要自定义实现。而对于查询对象有很多种，比如 termquery、booleanquery、fuzzyquery

等。不同的查询对象处理不同查询方式，如 termquery 主要用于查询词全匹配，booleanquery 则用于封装多个查询对象的关系等。正由于多元化的查询对象，才体现了 Lucene 的功能强大。 Lucene 在分词组件也提供丰富的接口以及实现方法，Lucene 提供几个分词器，如空格分词器、停顿词分词器等。当然同样也可以自定义分词器，进而实现满足自定义需求的分词器。可见，Lucene 在分词组件方面封装着非常好，具有很好的耦合性等特性。

2.6.3 Lucene 索引

2.6.3.1 倒排索引

倒排索引其实是用于记录词项与文档的对应关系的存储方式。在根据词项进行检索文档时，这种数据结构能够快速定位到包含该词项的文档集合，由于它与文档对应词项相反，所以叫做倒排索引。从上述的定义可知，倒排索引包含词典与文档集合两部分。倒排索引在搜索引擎中应用最为广泛，只要涉及到数据量很大时，并需要数据检索功能，倒排文件都是最好的选择。倒排文件也有很多变种，例如自索引[54]等。

2.6.3.2 Lucene 索引结构

该索引文件结构采用传统倒排文件来实现的，主要由索引库、索引段以及索引文档等三个部分组成，其中索引段又分为多个单一段索引。它们之间存在层次关系，即索引文档存储在单一段索引中，而索引文档中存储着域信息以及索引项[55]。Lucene 索引文件的概念组成和结构组成如图 2-8 所示。

图片

如图 2-8 所示，索引文件是由多个段索引组成的，而每个段索引则又由多个单一段索引组成，最后每个单一段索引存储着文档对象，并且每个文档对象存储着多个域对象，每个域对象存储着数据。该数据存储着词项信息以及它所在的文档的位置信息、权重信息等。

2.7 本章小结

本章首先对垂直搜索引擎的基础框架以及技术特点进行了描述，接着对网络爬虫技术、中文分词技术、非关系型数据库 HBase 以及全文搜索框架 Lucene 分别进行了详细阐述。其中网络爬虫技术主要详细介绍了主题爬虫架构以及三种网络爬虫策略，并分析两种爬虫策略的优缺点。在中文分词技术中详细介绍了三类分词算法的实现机制。而在非关系型数据库 HBase 中对关系型数据库与非关系型数据库的表结构进行了对比，并描述了两者之间的优缺点。最后在全文搜索引擎 Lucene 框架中主要详细介绍了 Lucene 的系统结构、索引、检索等部分，同时索引部分中介绍了索引结构、索引创建等部分，而检索部分中介绍了检索过程以及评分机制等部分。

1. 系统需求分析与设计

本章旨在阐述了面向招聘领域的垂直搜索引擎的需求分析及设计。首先介绍了面向招聘领域垂直搜索引擎的系统总体需求分析，主要包括功能需求分析、性能需求分析和安全需求分析。紧接着对招聘领域的垂直搜索引擎的整体设计进行描述，其中主要包括架构设计、系统的数据流、工作流程以及数据库的设计等；然后详细介绍垂直搜索引擎各个功能模块的层次关系以及详细设计，主要包括网页爬虫模块、中文分词模块、索引模块以及检索模块的设计；最后对搜索引擎的性能设计与安全设计进行介绍。

3.1 需求分析

3.1.1功能需求分析

针对垂直搜索引擎的相关技术以及招聘领域信息特点和用户需求进行研究，确定了以下需求：

1. 前台功能需求

前台的检索功能需要实现对关键字的检索，在检索结果中应该能够查看到标题、内容预览和网站链接，点击网站链接应该能够进入到该网页。每条检索功能应该能够清晰地分隔开，当检索结果较多的时候应该能够对检索结果进行分页。

1. 后台功能需求

（1）可以实时而又精确的采集全国每天的互联网上的招聘领域数据信息的分

布式主题爬虫系统。

（2）融合招聘领域的专业词库信息的中文分词器

（3）结合同义词以及 Rocchio 相关反馈算法对全文检索框架 Lucene 默认检索的查询优化，同时加入更为丰富的排序方式，使其更好的符合招聘领域的排序需求。

（4）友好而简洁的后台管理界面

3.1.2性能需求分析

（1）对用户搜索操作进行快速响应。用户提交检索后得到结果返回的等待时

间，系统响应时间越短，对用户来说体验性更好，其不仅取决于搜索引擎的性能同时与网络带宽也有关系，一般响应时间要低于Ｉｓ。

（2）支持高并发访问，当存在大量用户并发操作时，搜索响应不能延迟。

（3）搜索结果需具有较高的召回率与查准率［４３］。召回率是指搜索引擎所返回的检索结果站数据库中全部相关结果的比率。查准率是指符合用户需求的结果占所有查询结果的比例，取值范围在０？１之间。理想情况是两者越接近１越好，但

是实际情况是两者成反比。

（4）系统可靠性。应该采用成熟、稳定、通用的技术，保证系统的可靠性，数据具备自各份能力，能够保证系统长期稳定运行，有较强的容错和系统恢复能力，并为原系统提供调用接口。

（5）具有系统运行日志、数据更新日志等日志。

3.1.3安全需求

（1）数据安全。系统数据的安全一直是一个很重要的问题，本系统采用多种安全策略，保证数据的安全：对数据必须进行加密处理保证数据的机密性；采用数据验证机制保证数据的一致性和完整性；远程数据传输过程中的数据经过了SSL加密；用户信息采用MD5加密算法加密，将加密后的密文存入数据库用户信息表中，防止用户信息从数据库端泄露。这样处理同时可以有效避免SQL注入攻击。

（2）防御XSS攻击。跨站脚本攻击(Cross Site Scripting），简称XSS攻击，

是一种非常常见的攻击手段，攻击者利用WEB系统中的漏洞，通过在WEB前端页面中嵌入一些html和JavaScript的带有攻击性质的代码来实现恶意目的，造成系统出现一些不可预测的情况，攻击者可以利用漏洞和复杂脚本进行非法操作。

（3）防御CRSF攻击。跨站请求伪造（Cross－Site request forgery），简称CSRF攻击伪装成合法用户，替代原来合法用户实现入侵攻击目的，造成用户损失。跨站请求伪造（CSRF攻击）由于攻击者使用的并不频繁，这就导致针对这一攻击的检测、反入侵工具以及防范较少，如此一来，跨站请求伪造CRSF攻击一旦成功，造成的损失往往更大，难以防范。系统采用一些措施防止CRSF攻击，如将cookie设置为Http Only，对Http请求增加token，服务器端进行验证等。

3.2系统的整体设计

根据 3.1 节中提出的总体需求，本节对招聘领域的垂直搜索引擎系统进行了整体设计，主要包括架构设计、数据流设计、工作流程设计和HBase 数据库设计。

3.2.1系统架构设计

招聘领域的垂直搜索引擎整体架构如图 3-2 所示。

招聘领域的垂直搜索引擎由展示层、应用层、数据层以及存储层组成。展示层即本系统提供给用户的查询界面。应用层则是垂直搜索引擎中最重要的部分，它负责所有数据的处理以及业务逻辑操作，它包含很多核心模块，如网页采集及信息抽取、中文分词、索引和检索等。而数据层则是二手房垂直搜索引擎的数据中心，它负责系统中数据的读写操作，它直接影响着本系统的数据读写性能。存储层主要负责数据的存储方式，并提供数据存储的服务。

//图片

下面主要对应用层模块进行阐述。

（1）网页采集模块

本文采用域名预选策略并结合链接与主题相似度以及网页与主题相似度的判定，就可以很好的保证主题相关性。同时，为了能够不断的丰富可采集的资源，加入了对各大二手房网站平台的相关页面进行更新检测，如果检测到存在更新的二手房信息，则会通知爬虫服务进行抓取，同时，原本的爬虫服务会每隔一段时间循环抓取，从而实现对原网页库的更新，这样就实现了数据的全面性。 该设计方案可以保证数据的准确性和完整性，又因为主题判定时是基于文档结构树以及向量空间模型来实现的，复杂度不高，保证了爬虫的高效性。

（2）中文分词模块

中文分词模块主要是对网页采集模块提取的结构化二手房数据进行分词操作，本文采用目前比较流行的中文分词器 IKAnalyzer 来实现的，其中该分词器是基于正向迭代最细度切分算法实现的，并结合二手房领域专业词汇，使其能够有效的对二手房领域的文本进行分词。最后将其与全文检索框架 Lucene 进行整合应用到该系统中。

（3）索引模块

索引模块是基于全文检索框架 Lucene 来实现的。定时读取 HBase 数据库中二手房数据，为每个字段设置不同索引、分词等标志并存储到不同域中，追加更新到索引库中，以供后续检索模块的查询操作。

（4）检索模块

检索模块同样也是基于全文检索框架 Lucene 来实现的。利用全文检索框架Lucene 的特定项查询、布尔逻辑表达、模糊匹配、高亮查询等功能，实现了二手房特定字段查询、二手房全文信息检索等查询方式，同时整合全文检索框架Lucene 中默认检索模型与 Rocchio 算法来优化查询功能，使其检索结果更符合用户的需求，而且本系统还加入了特定字段的排序功能，从而延伸出更为个性化的排序功能，使购房者更为方便地找到满意的房源信息。

3.2.2系统数据流设计

本文招聘领域的垂直搜索引擎主要功能部分数据流图如图 3-3 所示。

//图片

如图 3-3 所示，数据流主要分为网页采集和信息检索两大部分，首先网页数据通过网页爬虫服务抓取下来，经过一系列的处理以及分析，将与主题相关的网页存储到原始网页数据库中，其中，中间处理操作包括 URL 相关性分析、页面内容相关性分析等，然后接着通过网页解析器，将非结构化数据解析成结构化数据，之后进行数据清洗操作，将重复信息以及异常数据剔除掉，得到清洗后的结构化数据，再对其建立索引库，最后实现好的检索器将根据用户输入的查询关键字，查询出所需的结果集返回给用户，这样就完成了数据在整个系统的处理以及变换流程。

3.2.3 分布式数据库 HBase 架构以及表设计

本文招聘领域的垂直搜索引擎采用 1 台主服务器与 3 台从服务器的设计方案来搭建分布式数据库 HBase 集群，而该集群又是基于 Hadoop 集群之上的，集群架构如图 3-4 所示。

//图片

如图 3-4 所示，分布式数据库 HBase 集群建立在 Hadoop 集群之上，通过ZooKeeper 集群服务来协调 1 台主服务器以及 3 台从服务器的之间的通信，从而完成对海量数据进行查询和分析等操作。招聘领域的垂直搜索引擎的分布式数据库 HBase 的表设计，包括行键设计以及列族的设计，这两者都会影响分布式数据库 HBase 的存储以及查询性能[77]，在本文的二手房的垂直搜索引擎中，每天又存在海量的二手房数据存储以及

查询操作，所以这两者的设计尤为重要，尤其是行键的设计，因为分布式数据库HBase 是根据行键来进行检索的，系统通过某个行键（或者某个行键范围）所在的分区，然后将查询数据的请求路由到该分区获取数据。而分布式数据库 HBase的行键以及列族的设计都有需要遵循的一定的规则，行键的设计规范如下。

（1）结合需求以及硬件等条件进行行键的设置。尽量让行键的长度保证在字节的整数倍，例如 8 字节大小。

（2）行键的设计不宜过长，过长的行键会导致内存的存储增大，并会影响检索操作的效率。

（3）行键的设计不宜采用连续增长的数字，因为这样的设计会导致数据存储到同一台节点服务器中，不利于节点服务器的负载均衡。而同样在列族的设计规则中，合理设计规范如下。

（1）表中列族的个数不宜超过 3 个。一个分区中存储着多个列族，而当一个列族中操作大量数据的时候，会引发文件写入操作，而这个操作会影响到其他的列族，同样文件的合并操作也会影响其他列族，所以这会引起多余的文件读写等步骤。

（2）由于每个数据文件都会存储列族以及列名等信息，所以必须保证名字简短些。并且，需要操作频繁的列名与不频繁的列名分开存储。 根据上述所说的规则，对二手房数据进行行键及列族设计，并结合分布式数据库 HBase 的特性，确定出二手房数据的表结构的设计。 二手房数据是通过网页采集模块采集得到的原始网页数据库，接着经过网页解析器得到的结构化数据信息，其中结构化数据各属性描述如表 3-1 所示。

//表

如表 3-1 所示，这里标识了二手房结构化数据各个属性以及描述，接下来详细介绍该结构化数据表的设计。根据结构化数据信息、系统的功能需求以及性能需求分析、分布式数据库HBase 的行键设计规则等综合因素，对其进行行键的设计如表 3-2 所示。

//表

如表 3-2 所示，行键分为 8 个字节，刚好对齐计算机 64 位操作系统，内存8 字节，这样利于操作系统的存取数据的速度，在第 0 到 3 个字节存放当前整点时间格式的散列值，其中当前整点时间指的是当前日期加上当前小时的组合，它保证每个小时爬取的数据存放在不同的节点服务器中，这样可以在每个节点服务器实现负载均衡，同时也是为了提高后续的时间段的查询效率，让其追加到索引库中，从而实现实时检索二手房信息。由于该段数据是一段连续时间的数据，所以这样的设计是有利于提高查询效率的。其余 4 个字节用于唯一标识，可以解决行键重复的问题。在完成行键的设计之后，紧接着列族的设计，同样按照列族的遵循的设计规则，需要将字段分配到不同的列族当中，不可能将所有的字段信息都设计到一个列族里，所以列族的设计如表 3-3 下。

//表

如表 3-3 所示，表中包含两个列族，即 MInfo 和 EInfo，这样的设计是出于列族 MInfo 主要是存放用于构建索引的字段，这些字段会在后续的操作中被查询，所以才放置到同一个列族当中，而对于列族 EInfo主要存储暂时不需要被查询的字段，这样的设计即不影响后续的索引构建，同样还有提供很好的查询效率。另外设计成两个列族也满足不超过 3 个列族的规则。

以上就是分布式数据库 HBase 中表的设计，在本文招聘领域的垂直搜索引擎系统中通过采用以上的设计实现了数据在每个节点服务器上的负载均衡，同时也实现了系统对查询性能的要求。

3.3网页爬虫模块的设计与实现

招聘领域的垂直搜索引擎的网页爬虫模型是基于一个轻量级的分布式爬虫框架 JLiteSpider 以及 HBase 数据库集群来实现的，轻量级的分布式爬虫框架JLiteSpider 主要负责业务逻辑层的实现，而 HBase 数据库集群负责数据库的存储工作。

3.3.1 分布式爬虫框架 JLiteSpider

该分布式爬虫框架 JLiteSpider 是基于 java 实现的，它是一个强大，但又轻量级的分布式爬虫框架。该爬虫框架分布式的实现方式是基于任务以及消息队列来实现的，各个任务之间通过一个或者多个消息队列来链接，这里消息队列采用Rabbitmq 来实现的；任务与消息之间可以是一对一、一对多、多对一以及多对多的关系。消息队列中存储的消息分为四种：URL、页面源码、解析后的结果以及自定义的消息。同样的，任务的工作也分为四部分：下载页面、解析页面、数据持久化以及自定义的操作。 用户只需要在配置文件中，规定好的任务和消息队列之间的关系。接着在代码中，定义好任务的四部分工作机制。即可完成分布式爬虫的任务。 该分布式爬虫框架 JLiteSpider 的使用流程如下：

（1）启动 Rabbitmq 服务。

（2）在配置文件中定义任务以及消息队列的关联

（3）编写相应任务的业务逻辑

（4）最后，启动爬虫服务

5.3.2 爬虫服务架构设计

本文针对网站的招聘信息发布情况、各个招聘网站的网页布局、更新情况以及详情页面的结构等因素，同时结合对国内网站的招聘信息真实性的调研，选取出符合招聘信息的抓取的网站，如智联招聘、前程无忧、拉勾网、BOOS直聘等网站。本文通过对这些网站中的招聘数据信息进行抓取，从而构建招聘网页数据源。 由于所抓取的网站数量较多，本文对不同的网站构建相对应的分布式网页爬虫服务引擎，这样可以更好的抓取到准确以及全面的招聘数据信息。每个分布式网页爬虫服务引擎负责该网站的招聘信息抓取工作，接下来对“前程无忧”网站的分布式爬虫服务引擎架构图进行详细介绍。具体架构图如图 3-5 所示。

//图

如图 3-5 所示，分布式爬虫引擎 JLiteSpider 从房天下 URL 种子库中，获取到初始 URL 链接，进而开始网页爬虫服务，分布式爬虫引擎 JLiteSpider 包含多个工作单元协同工作，每个工作单元中包含下载页面模块、解析网页模块以及数据持久化模型等。下载页面模块从待访问 URL 库中获取与主题相关网页链接，通过 Http 请求将网页源码下载下来，利用 Rabbitmq 消息队列引擎将网页源码传递给解析网页模块，解析网页模块分析出与主题相关的网页链接存放到待访问的URL 库，同时也会将源码传递给数据持久化模块，进而数据持久化模块通过调用 HBase 数据层中间件，将网页数据信息存储到 HBase 数据库集群中，进而完成对“前程无忧”网站中招聘信息的分布式爬虫服务。

5.3.3 网页抓取

本文将抓取数个网站中招聘信息页面，根据前面所提到的设计，每个分布式爬虫服务引擎针对一个网站所有城市的二手房页面进行抓取。这里以“前程无忧”网站为例。首先将网站的每个城市招聘信息模块链接地址添加到 URL 种子库中，启动分布式爬虫服务引擎，至于分布式爬虫服务中任务数可以根据网站的 URL 种子数来定，爬虫服务引擎启动之后，下载页面模块会请求访问地址，将网页下载下来，然后根据链接相似度计算模块并结合文档解析模块将与二手房主题相关的链接提取出来，并加入待访问 URL 队列中，同时也会根据网页相似度计算模块来保存原始网页，并解析出结构化数据与其一同存储到 HBase 数据库集群中。这里设置起始 URL 为 http://esf.cd.fang.com，后台爬虫系统如图 3-6 所示。

//图

3.3.4 文档解析与处理模块

网页文档是由 html 标签组织成的非结构化的文本信息，对于垂直搜索引擎而言，查询的结果是一个结构化数据信息，所以如何将招聘信息中的关键属性信息提取出来成为了本节所要解决的问题。

3.3.4.1 网页解析工具 HTMLparser

本系统利用 HTMLparser 对二手房网页进行结构化数据的提取、预处理以及分析等操作，HTMLparser 提供了一系列的网页解析方法。HTMLparser 是纯 Java编写的开源轻量级网页解析工具库[78]，通过将网页解析成文档树，进而对网页各个节点进行遍历、查找、删除等操作。它可以快速的提取各个标签元素的文本以及属性信息。HTMLparser 在解析网页方面有着速度快的特点。 HTMLparser 主要提供对文档中关键信息的提取以及对提取之后的信息变换等功能。

（1）信息提取主要是对网页中的文本信息、链接以及图片等信息的提取，当然除了信息的提取，HTMLparser 还提供辅助功能来补充信息提取功能。例如对图片、声音等资源信息得检查以及分析，还有对网页中的链接进行检查，确定该链接是否为有效的网页链接等。

（2）为了对提取好的信息进行分析、检查等处理，HTMLparser 提供了一系列的信息转换的接口，主要包括对链接的重写、网页内容检验等方法。同时还可以利用正则表达式对内容信息的过滤，例如垃圾信息、广告等噪声信息。

5.3.4.2 信息抽取

网页文档解析是从网页下载器中获得网页源代码，然后结合 HTMLparser 组件对网页文档构建文档结构树，为要抓取的二手房的属性信息设置不同的 Xpath解析规则，然后结合文档结构树解析出二手房的各个属性信息，即完成了信息提取过程。HTMLparser 组件对信息提取有着很高的效率，并且它的容错率也很高，同时也是目前最好的解析网页的工具库。本文采用 HTMLparser 进行网页解析与信息抽取，流程图如图 5-9 所示。

网页解析与信息抽取过程：

（1）对整个网页源码进行文档结构树构建，为不同属性定制不同的 Xpath解析规则，根据此规则可以提取出包含标题、价格、户型、楼盘、楼层等相关属性信息。

（2）遍历文档结构树中的节点对象，提取符合各个 Xpath 解析规则的结构化数据。

（3）对解析好的结构化数据存储到 HBase 数据库中。

//图

采用 HTMLparser 对抓下来的 HTML网页文档进行网页解析与信息抽取后，将结构化的数据保存到 HBase 数据库集群中。各个属性的解析 Xpath 的格式如表 5-4 所示，以及解析出的数据存储到 HBase 集群中的数据格式如表 5-5 所示，这里并没有将所有列名的数据表示出来，仅仅列出了“价格”、“朝向”、“城区”、“楼盘名”等列名数据。

//表

3.4 索引模块的设计与实现

在搜索引擎中，索引设计的好坏会直接影响搜索结果的准确性，如果对那些无关的数据进行索引，会导致用户搜索不到相关的数据信息，同时还会导致索引文件膨胀等问题，同样分词的好坏也会影响搜索引擎的检索效果，索引模块的设计包括对数据预处理、中文分词器、索引构建等。

5.4.1 数据预处理

数据预处理主要是对结构化的招聘信息数据去重等操作，因为同一个招聘信息有可能在不同的网站上发布过，这样导致招聘信息存在重复问题。本文通过短文本相似度来计算招聘信息的相似度，进而实现招聘信息的去重功能。由于招聘信息的数据量很大，原始数据库中存放在上千万条数据，如果全表扫描数据库的招聘信息数据，然后依次进行相似度计算，这样的设计复杂度会很高，同时结合根据网站信息的招聘信息发布规律，本文设定一个月的招聘数据来进行相似度计算，从而降低相似度的复杂度，同时也能够降低招聘信息的重复率。 短文本相似度算法有很多种，如欧式距离、余弦距离、编辑距离、Jaccard 相似度、Shingling 相似度等，本文采用结合 Shingling 与 Jaccard 相似度来实现短文本相似度的计算。

（1）Jaccard 相似度

Jaccard 相似度主要是用在比较两个集合的相似度，将两个集合的交集与并集的比例作为两个集合的相似度。

（2）Shingling 算法

Shingling 算法可以整合 Jaccard 相似度算法来实现文档之间的相似度计算，通过将文档转换为连续 k 字符的集合，然后结合 Jaccard 相似度来求出文档之间的相似度。本文利用对文档中每 2 个连续的字符进行散列，得到一组散列值组，用该散列值集合表示文档更为有效。

（3）编辑距离算法

该算法是通过统计两个字符串之间转换成对方的步骤，将转换的次数作为两个字符串之间的相似度，其中转换的操作包含插入、删除两个操作[79]。

（4）海明距离

海明距离主要通过统计两个向量之间各个位置不同的个数，将其作为两个向量的距离值。 假设两个短文本分别是 S1和 S2,即 S1=“成都季柳园 2 室 2 厅 100 万二手

房”，S2=“成都交大季柳园 2 室 2 厅二手房”，设定 k=2，计算 S1和 S2的相似度。 S1 和 S2相似度计算过程是对 S1和 S2中的每连续 2 个字符分割，然后对连续字符进行散列得到散列码，并将其保存到数组中，得到两个存放散列码的数组S1\_arr 和 S2\_arr，具体数组如下。 S1arr= [815341, 1174022, 751856, 847194, 689893, 25010, 727310, 22931, 662860, 1567, 1536, 21463, 639333, 648511] S2\_arr= [815341, 1170759, 646915, 730908, 751856, 847194, 689893, 25010, 727310, 22931, 682919, 648511] 接着对 S1\_arr和 S2\_arr集合中取交集 c，然后利用 Jaccard 相似度计算方式，则S1和 S2相似度为 sim(S1,S2)=c/max(|S1\_arr|,|S2\_arr|)=8/14≈0.5174。

3.4.2 IKAnalyzer 中文分词器

中文分词算法在第 2 章已经介绍了，目前国内有很多的中文分词工具都是基于中文分词算法来实现的，本文采用 IKAnalyzer 分词工具作为本系统实现中文分词模块的基础，IKAnalyzer 是一个开源的，基于 java 语言开发的轻量级的中文分词工具包。本文整合 IKAnalyzer 工具类到搜索引擎系统的流程如图 5-10 所示。

//图片

本系统根据爬虫服务搜集二手房领域专业词库以及各个城市楼盘名等数据，构建二手房领域词库用作扩展词库，同时结合基础词典，在全文检索框架 Lucene基础上整合 IKAnzlyzer 中文分词器实现中文分词模块，这里举例说明。 假设现在对二手房文档集合进行分词处理，分词效果如表 5-6 所示。

//表

3.4.3 索引构建

本文对二手房结构化数据进行索引构建，整体流程如下所示，其中，索引构建流程如图 5-11 所示。

（1）配置二手房领域的词典并创建中文分词器 IKAnalyzer 对象。

（2）设置索引创建目录与分词器并创建写索引对象 IndexWriter。

（3）创建 HBase 数据库读取对象，并读取整点时间段的二手房数据信息。

（4）对整点时间段的二手房数据信息去重，采用 5.4.1 所提到的短文本相似度计算，将重复二手房的数据删除。

（5）对检索出来的二手房数据进行添加到索引库中，如果没有添加完成，则依次继续完成每条二手房数据的索引添加操作。如图 5-11 所示。

（6）如果整点时间段的二手房数据完成添加，则 IndexWriter 进行提交。

（7）关闭所有连接对象。

在索引构建过程中，还有对索引的更新操作，本系统通过设置定时任务，判定 HBase 数据库中是否存在新的二手房数据，存在则会进行索引的追加操作，操作流程与该流程一致。

//图

3.5 检索模块的设计与实现

3.5.1 检索过程

信息检索的流程是通过给定一个用户输入的查询词，对查询词进行中文分词以及查询扩展，接着对索引库中词项进行检索，得到所有具有评分的文档集合并结合 Top-k 排序算法将前 k 个评分最大的结果集返回给用户，而其中具体评分方式由排序方式来决定，例如相关性排序、特定域排序等。本文设计的二手房垂直搜索引擎对检索模块中两个检索方式都进行了实现，即默认查询与扩展查询，默认查询采用的是通用全文检索框架 Lucene 的检索方式，而扩展查询采用的是第四章提出的融合同义词和 Rocchio 的相关反馈算法来优化查询。检索模块的流程

如图 5-12 所示。

//图

具体检索过程如下所示。

（1）用户输入查询词、查询方式、查询域。

（2）指定索引目录并打开索引读取对象 IndexReader。

（3）创建文档检索实例 IndexSearcher 并设置相似度计算类型。

（4）利用 IKAnalyzer 分词器对查询词进行分词操作。

（5）判断是否对查询向量进行优化，如果查询方式为“扩展查询”，则根

据同义词优化与 Rocchio 相关反馈算法，优化查询向量。

（6）通过设置域信息以查询向量构建查询对象 Query。

（7）根据查询对象 Query 进行检索，将检索出的文档集合按照指定排序算

法以及 k 的值，将 k 个评分的查询结果集返回浏览器。

5.6 本章小结

本章旨在对二手房的垂直搜索引擎的设计与实进行详细的介绍。首先是对

二手房的垂直搜索引擎的需求进行分析，确定该系统的功能以及目标。接着对垂

直搜索引擎中的数据库架构设计以及表的设计，主要包括 HBase 数据库集群的

设计以及房源表的设计，其中房源表的设计涉及到行键的设计以及列族的设计。

接着对网页爬虫模块的设计与实现进行了介绍，其中包括基于分布式爬虫框架

JLiteSpider 实现一套抓取房产网站中二手房模块的爬虫框架服务程序以及网页

解析模块。然后对索引模块的设计与实现进行了介绍，主要包括数据预处理、中

文分块以及索引构建等部分。最后对检索模块进行详细的介绍，主要包括检索过

程以及查询效果等部分。

# 参考文献

[1] 计算广告[M]. 人民邮电出版社 , 刘鹏, 2015

[2] 统计学习方法[M]. 清华大学出版社 , 李航, 2012

[3] 广告点击率预估技术综述[J]. 浙江理工大学学报, 陈巧红,余仕敏,贾宇波. 2015(11).

[4] Logistic回归模型[M]. 高等教育出版社 , 王济川,郭志刚[著], 2001

[5] [基于大数据平台的精准广告系统研究与设计](https://kns.cnki.net/kcms/detail/detail.aspx?filename=DNXJ201504015&dbcode=CJFQ&dbname=CJFD2015&v=)[J]. 张建,孙铭,段娟.  电脑与信息技术. 2015(04)

[6] [计算广告:以数据为核心的Web综合应用](https://kns.cnki.net/kcms/detail/detail.aspx?filename=JSJX201110007&dbcode=CJFQ&dbname=CJFD2011&v=)[J]. 周傲英,周敏奇,宫学庆.  计算机学报. 2011(10)

[7] [电子商务推荐模型的用户特征提取方法及系统](https://kns.cnki.net/kcms/detail/detail.aspx?filename=CN104331817A&dbcode=SCPD&v=)[P]. 沈慧,范小朋,赵东辉,须成忠. 中国专利:CN104331817A,

[8] [大数据时代背景下的广告营销模式研究](https://kns.cnki.net/kcms/detail/detail.aspx?filename=1014297322.nh&dbcode=CMFD&dbname=CMFD2014&v=)[D]. 张之怡.苏州大学 2014

[9] [面向自然语言处理的深度学习研究](https://kns.cnki.net/kcms/detail/detail.aspx?filename=MOTO201610001&dbcode=CJFQ&dbname=CJFD2016&v=)[J]. 奚雪峰,周国栋.  自动化学报. 2016(10)

[10] [在线广告中高层特征表示及点击率预测方法研究](https://kns.cnki.net/kcms/detail/detail.aspx?filename=1014083837.nh&dbcode=CMFD&dbname=CMFD2015&v=)[D]. 邵迪.哈尔滨工业大学 2014

[11] [不同池化模型的卷积神经网络学习性能研究](https://kns.cnki.net/kcms/detail/detail.aspx?filename=ZGTB201609007&dbcode=CJFQ&dbname=CJFD2016&v=)[J]. 刘万军,梁雪剑,曲海成.  中国图象图形学报. 2016(09)

[12] [广告点击率预估的深层神经网络模型研究](https://kns.cnki.net/kcms/detail/detail.aspx?filename=1015583935.nh&dbcode=CMFD&dbname=CMFD2015&v=)[D]. 王孝舒.北京邮电大学 2015

[13] [基于特征学习的广告点击率预估技术研究](https://kns.cnki.net/kcms/detail/detail.aspx?filename=1017241252.nh&dbcode=CMFD&dbname=CMFD2018&v=)[D]. 周永.哈尔滨工程大学 2014

[14] [基于特征学习的广告点击率预估技术研究](https://kns.cnki.net/kcms/detail/detail.aspx?filename=JSJX201604010&dbcode=CJFQ&dbname=CJFD2016&v=)[J]. 张志强,周永,谢晓芹,潘海为.  计算机学报. 2016(04)

[15] [基于LDA的互联网广告点击率预测研究](https://kns.cnki.net/kcms/detail/detail.aspx?filename=JSYJ201604004&dbcode=CJFQ&dbname=CJFD2016&v=)[J]. 朱志北,李斌,刘学军,胡平.  计算机应用研究. 2016(04)

[16] 计算广告学将成为数字商业的奠基学科[J]. 李敏.  程序员. 2014 (05)

[17] [基于平衡采样的轻量级广告点击率预估方法](https://kns.cnki.net/kcms/detail/detail.aspx?filename=JSYJ201401007&dbcode=CJFQ&dbname=CJFD2014&v=)[J]. 施梦圜,顾津吉.  计算机应用研究. 2014(01)

[18] [广告点击率估算技术综述](https://kns.cnki.net/kcms/detail/detail.aspx?filename=HDSZ201303005&dbcode=CJFQ&dbname=CJFD2013&v=)[J]. 纪文迪,王晓玲,周傲英.  华东师范大学学报(自然科学版). 2013(03)

[19] [基于联合概率矩阵分解的上下文广告推荐算法](https://kns.cnki.net/kcms/detail/detail.aspx?filename=RJXB201303002&dbcode=CJFQ&dbname=CJFD2013&v=)[J]. 涂丹丹,舒承椿,余海燕.  软件学报. 2013(03)

[20] [基于贝叶斯网的广告点击率预测方法及实现](https://kns.cnki.net/kcms/detail/detail.aspx?filename=1013306802.nh&dbcode=CMFD&dbname=CMFD2014&v=)[D]. 王朝禄.云南大学 2013

[21] [网络广告投放算法的研究](https://kns.cnki.net/kcms/detail/detail.aspx?filename=1014183089.nh&dbcode=CMFD&dbname=CMFD2014&v=)[D]. 霍艳.东北大学 2013

[22] [大数据分析的无限深度神经网络方法](https://kns.cnki.net/kcms/detail/detail.aspx?filename=JFYZ201601007&dbcode=CJFQ&dbname=CJFD2016&v=)[J]. 张蕾,章毅.  计算机研究与发展. 2016(01)

[23] [面向大数据分析的在线学习算法综述](https://kns.cnki.net/kcms/detail/detail.aspx?filename=JFYZ201508002&dbcode=CJFQ&dbname=CJFD2015&v=)[J]. 李志杰,李元香,王峰,何国良,匡立.  计算机研究与发展. 2015(08)

[24] 数据挖掘中的新方法[M]. 科学出版社 , 邓乃扬,田英杰著, 2004

[25] [协同过滤推荐算法综述](https://kns.cnki.net/kcms/detail/detail.aspx?filename=XXWX200907006&dbcode=CJFQ&dbname=CJFD2009&v=)[J]. 马宏伟,张光卫,李鹏.  小型微型计算机系统. 2009(07)

[26] [一种基于逻辑回归模型的搜索广告点击率预估方法的研究](https://kns.cnki.net/kcms/detail/detail.aspx?filename=1013178340.nh&dbcode=CMFD&dbname=CMFD2014&v=)[D]. 王兵.浙江大学 2013

[27] [个性化广告点击率预测的研究和实现](https://kns.cnki.net/kcms/detail/detail.aspx?filename=1013241851.nh&dbcode=CMFD&dbname=CMFD2013&v=)[D]. 司向辉.北京邮电大学 2013

[28] Ying Shan, T Ryan Hoens, Jian Jiao, Haijing Wang, Dong Yu, and JC Mao. 2016. Deep Crossing: Web-Scale Modeling without Manually Cra ed Combinatorial Features. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 255–262.

[29] Wang, R., Fu, B., Fu, G., Wang, M.: Deep & cross network for ad click predictions. In: Proceedings of the ADKDD 17. pp. 12:1–12:7 (2017).

[30][Gai et al] Learning Piece-wise Linear Models from Large Scale Data for Ad Click Prediction

[31][Cheng et al., 2016] Wide & deep learning for recommender systems. CoRR.

[32][Huifeng Guo et al, 2017] DeepFM: A Factorization-Machine based Neural Network for CTR Prediction.

[33][Guorui Zhou et al, 2017] Deep Interest Network for Click-Through Rate Prediction.

[34][He X, Pan J, Jin O, et al, 2014] Practical lessons from predicting clicks on ads. at facebook. ACM SIGKDD.

[35][Rendle, 2010] Factorization machines. In ICDM.

[36] Practical Lessons from Predicting Clicks on Ads at Facebook. He,X.,et al. ADKDD’’14 . 2014

[37] Paul Covington, Jay Adams, and Emre Sargin. Deep neural networks for youtube recommendations. In Proceedings of the 10th ACM Conference on Recommender Systems, pages 191–198. ACM, 2016.

[38] Heng-Tze Cheng and Levent Koc. Wide & deep learning for recommender systems. In Proceedings of the 1st Workshop on Deep Learning for Recommender Systems, pages 7–10. ACM, 2016.

[39] Scalable training of L1-regularized log-linear models. Andrew G,Gao J. Pro-ceedings of the24th International Conference on Machine Learning . 2007

[40] Dual Averaging Methods for Regularized Stochastic Learning and OnlineOptimization. Lin X. NIPS . 2009

[41] [[GBDT+LR] Practical Lessons from Predicting Clicks on Ads at Facebook (Facebook 2014)](https://github.com/wzhe06/Ad-papers/blob/master/Classic%20CTR%20Prediction/%5BGBDT%2BLR%5D%20Practical%20Lessons%20from%20Predicting%20Clicks%20on%20Ads%20at%20Facebook%20%28Facebook%202014%29.pdf)

[42] [XGBoost：A Scalable Tree Boosting System](https://arxiv.org/pdf/1603.02754.pdf)