单位代码：10166

沈阳师范大学

硕士学位论文

自动答疑系统的设计与实现

论文作者：马新意

学科专业：计算机应用技术

指导教师：王剑辉

培养单位：数学与系统科学学院

培养类别：全日制

完成时间：2019 年 05 月 18 日

沈阳师范大学学位评定委员会

学位论文独创性声明

本人所呈交的学位论文是在导师的指导下取得的研究成果。据我所知，除文中已经注明引用的内容外，本论文不包含其他个人已经发表或撰写过的研究成果。对本文的研究做出重要贡献的个人和集体，均已在文中作了明确说明并表示了谢意。

作者签名：马新意 日期：2019.5.18

学位论文使用授权声明

本人授权沈阳师范大学研究生处，将本人硕士学位论文的全部或部分内容编入有关数据库进行检索；有权保留学位论文并向国家主管部门或其指定机构送交论文的电子版和纸质版，允许论文被查阅和借阅；有权可以采用影印、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适用本规定。

作者签名：马新意 日期：2019.5.18

编　号：

类别		
	全日制研究生	√
	教育硕士	
	同等学力	

沈阳师范大学

硕士学位论文

题　目：　自动答疑系统的设计与实现

培　养　单　位：　数学与系统科学学院

专　业　名　称：　计算机应用技术

指　导　教　师：　王剑辉

研　　究　　生：　马新意

完　成　时　间：　2019 年 5 月 18 日

沈阳师范大学研究生处制

自动答疑系统的设计与实现

中文摘要

随着教学资源的完善和网络技术的发展，教学的中心从过去的以教师的教不断转变为基于学生独立探究的新教学模式，在学生主观能动性的促使下，他们发现问题、提出问题、探索问题、解决问题，汲取知识的热情达到了空前的高度，如何有效率的解决学生在课堂课后遇到的问题，成为越来越多教师的难题，至此自动答疑系统应运而生。自动答疑系统是中文信息检索领域的一个重要应用，作为课堂的补充和延续，学生与系统进行交互时，它可以消除学生紧张、害羞和恐惧的心理，以轻松、自由、积极的态度找寻问题的答案，这对学生的心理发展和自主学习起到了至关重要的作用。自动答疑系统有效地减少了教师的工作量，将教师从繁重、重复的答疑工作中解脱出来，将工作重心放在提高学生学习效率方面，同时促进了学生独立思考和个性化学习。此外，它可以促进在线网络教学的发展，通过对问题记录的分析，教师可以发现学生的一些薄弱环节，为改进教学方法，制定新的教学策略提供参考。

本文比对了国内外自动答疑系统的研究现状，概述了Lucene的系统结构和运行机制，并将Lucene全文搜索引擎、SQL like子句、MYSQL数据库三种方式进行检索性能的比较，选择Lucene全文搜索引擎作为自动答疑系统的检索工具。Lucene全文搜索引擎在具有强大检索能力的同时也存在着致命的弱点，由于Lucene中没有完善的中分分词模块，且系统自身的中文分词法的分词准确性达不到答疑的标准，本文在参考了分词词典机制的基础上，比较了三种分词机制的优缺点，提出了一种基于双字哈希索引的词典机制。基于使用范围、操作难易等多方面的考量，选择正向最大匹配算法作为系统的中文分词算法，并对该算法进行了改进，改进后的分词法不再需要人为设定最大词长，能够在运行的过程中自动获取最大词长。将改进的分词算法应用于Lucene系统架构中，与统计法相结合使得它对交叉型歧义具有一定的消歧效果。根据自动答疑系统的特点，提出了利用自然语言处理技术，支持特定科目问题的系统模型，分析和讨论了关键技术，对其中重要功能模块进行设计和实现。系统理解用户输入的自然语言，并返回具有最高相似度的答案。经实验证明，自动答疑系统具有良好的切分精度和实用性。

关键词：自动答疑系统，Lucene 全文检索，最大匹配算法，双字哈希索引

Design and implementation of automatic question answering system

Abstract

With the improvement of teaching resources and the development of network technology, the center of teaching has changed from teachers' teaching to a new teaching mode based on students' independent inquiry. Under the impetus of students' subjective initiative, they find problems and put forward problems. The enthusiasm for exploring, solving and absorbing knowledge has reached an unprecedented height. How to solve the problems encountered by students after class efficiently has become a difficult problem for more and more teachers, and the automatic question answering system emerges as the times require. Automatic question answering system is an important application in the field of Chinese information retrieval. As a supplement and continuation of the classroom, it can be eliminated when students interact with the system. Students are nervous, shy and afraid, and look for answers to questions in a relaxed, free and positive manner, which plays an important role in students' psychological development and autonomous learning. The automatic question answering system effectively reduces the workload of teachers, frees teachers from the heavy and repetitive answer work, and focuses on improving students' learning efficiency, at the same time, it promotes students' independent thinking and individualized learning. In addition, it can promote the development of online network teaching. Through the analysis of problem records, teachers can find some weak links of students, and provide reference for improving teaching methods and formulating new teaching strategies.

This paper compares the research status of automatic question answering system at home and abroad, summarizes the system structure and running mechanism of Lucene, and compares the retrieval performance of Lucene full-text search engine, SQL like clause and MYSQL database. Lucene full-text search engine is selected as the retrieval tool of automatic question answering system. Lucene full-text search engine not only has strong retrieval ability, but also has fatal weakness, because there is no perfect middle segmentation module in Lucene. And the accuracy of Chinese word segmentation in the system itself can not meet the standard of answering questions. In this paper, the accuracy of word segmentation in Chinese is not up to the standard of answering questions. On the basis of referring to the word segmentation dictionary mechanism, the advantages and disadvantages of the three word segmentation mechanisms are compared, and a dictionary mechanism based on double character hash index is proposed. Based on the consideration

of the scope of use, the difficulty of operation and so on, the forward maximum matching algorithm is selected as the Chinese word segmentation algorithm of the system, and the algorithm is improved. The improved word segmentation method no longer needs to set the maximum word length artificially. Can automatically get the maximum word length in the process of running. The improved word segmentation algorithm is applied to the architecture of Lucene system, which is combined with the statistical method to make it have a certain effect on cross ambiguity. According to the characteristics of automatic question answering system, this paper puts forward the method of automatic question answering. This paper makes use of natural language processing technology to support the system model of specific subject problem, analyzes and discusses the key technology, and designs and implements the important functional modules. The system understands the natural language entered by the user and returns the answer with the highest similarity. It is proved by experiments that the automatic answer system has good precision and practicability.

Key words:
Automatic question answering system, Lucene full-text search, maximum matching algorithm, double-character hash index

目录

第一章 引言

随着远程教学的普及和不断完善，越来越多的学者选择基于网络的在线教学模式。在线网络教学突破了传统教学对年龄、时间和空间方面的限制，无论你是谁，无论你身处何地，随时都可以打开手机或者电脑进行在线学习。网络在线学习在具有诸多优势的同时，也存在着一个弊端，那就是如何解决学者在学习过程中遇到的疑难问题呢，在这种情况促使下，出现一个能够为学习者答疑解惑的自动答疑系统是相当必要的。

一、研究背景

如今，随着教学资源的完善和网络技术的发展，教学的中心从过去的以教师的教学不断转变为基于学生独立探究的新教学模式，在学生自主发现问题、提出问题、探究问题、解决问题的学习过程促使下，如何有效率的解决学生在课堂课后遇到的问题，成为越来越多教师的难题。在传统的教学中大多数教师采用的方法是开设专门的答疑课，在答疑课上学生以四人小组为单位进行答疑讨论，简单的问题组内解决，困难的问题以小组为单位向教师提问，教师整理全班的问题后统一讲解。开设答疑课的优势在于提升了学生自我探索的能力，弊端是学生对知识的理解具有相似性，大量重复性的问题消耗了教师的精力和课堂时间，答疑作为教学与学习的重要组成部分，正日益引起教育工作者的关注。

现在随着教育信息化进程的推进，越来越多的学生选择基于网络的自主性学习，通过网络学生可以进行个性化学习和课后辅助答疑。如今全国许多高等院校都使用网络教学平台，作为课堂教学的补充和延伸。网络教学平台在发布教学材料、评估学生学习情况、发布课程相关材料、布置课后作业的基础上，通过学生在答疑论坛中的交流，提高学生独立思考的能力，还有效减轻了教师的工作量，教师可以通过对提问记录的分析来发现学生普遍存在的薄弱环节，为改进教学方法、制定新的教学策略、部署教学事项提供了参考。

二、研究意义

国内的自动答疑系统大多嵌套于网络教学平台中，作为网络教学平台的一个功能模块，具有单独答疑功能的系统非常少。如今许多答疑系统从用户的提问中提取出关键词语，利用搜索引擎检索后，搜索引擎将与该词语相关的内容全部呈现给用户。首先这要求用户具有准确提取关键词的能力，更要掌握相关的逻辑知识，其次呈现给用户的并不是一个或几个确定的答案，而是一组缺乏相似度排序的材料，需要用户在海量的材料中进行二次检索以找到真正需要的答案，在检索的过程中更容易被错误的答案误导，消耗了大量的时间，

这会影响学生固有的知识体系。由于中文语义的复杂性以及中文分词技术的难度，国内的答疑系统存在着检索效率低、智能化程度不高、查全查准率不高、系统跨平台性不好等不足之处，要想完全实现自然语言交互并投入实际应用还有很长的路要走。但是具体到特定的科目，由于学生的提问是面向一门特定的课程，每门课程都有自己固定的知识体系结构，知识点间的关系由浅入深、循序渐进，不同地区的教材在内容上也不会有太大的区别，只是组织顺序上有所不同。考虑到学生的接受能力，他们的提问具有一定的相似性，只是在表达方式上略有不同。因此设计实现一个易交互、可行性高、操作简便的针对特定科目的自动答疑系统是可行的。

三、研究现状

（一）国外研究现状

国外大多数比较成功的智能答疑系统都是独立运行的系统，不依附于任何网络教学平台，有 AskJeevesforKids 网上智能答疑系统、START 答疑系统、AnswerBus 和 FAQFind 答疑系统等。从智能性来看，由于西方语句自带空格成为天然分隔符的特点，使得国外答疑系统的交互性更好，由于对分词技术的研究较早，使得系统有更高的查全率与查准率。

START[1](Syn Tactic Analysis using Reversible Transformation)答疑系统于 1993 年问世，自 1993 年 12 月以来一直在线并持续运行，它是最早出现的答疑系统，也是首个面向网络基于自然语言问答的系统，由 MIT 人工智能实验室 Infolab Group 的 Boris Katz 及其同事合作开发。其网址为：http://www.ai.mit.edu/projects/infolab。与信息检索系统和搜索引擎不同，Start 系统不仅根据浏览次数的多少提供点击量列表，其主要目的是通过交互向用户提供正确的信息。目前，该系统可以回答关于地理（如城市、国家、湖泊、坐标、政治）、电影（如演员、导演、头衔）、人物等相关问题，问题的答案也不局限于文字，也可以与图片、声音、动画等内容组合呈现。

AskJeevesforKids 网上智能答疑系统由美国 AskJeeves 公司开发，其网址为：http://www.askieeves.com。AskJeevesforKids 系统允许用户以自然语言进行提问，并通过与用户的交互逐步确定用户的真正意图，提高检索的精确度。

AnswerBus[2]是 Zhiping Zheng 博士组织开发的答疑系统，它是基于搜索引擎的答疑系统，以 Google,Yahoo,Yahoo News,Alta Vista 和 Wisenut 作为搜索引擎，支持英语，德语，法语，西班牙语，意大利语和葡萄牙语等多种语言进行的自然语言查询系统。

FAQFind 答疑系统由芝加哥大学人工智能实验室开发，它预先构建了问答形式的 FAQ 库，通过语义分析，采用基于向量的检索方式在 FAQ 库中检索答案。

（二）国内研究现状

国内对答疑系统的研究始于20世纪90年代末，提出了多种理论并经历了几个重要阶段。从一开始没有单独的答疑部件，师生通过留言板、电子邮件等方式进行交流；到具有初步的答疑部件，如清华大学的远程教育系统，教师和学生既可以异步讨论，也可以进行WEB实时交流；然后逐步发展出具有自动答疑功能的答疑部件，如上海交通大学设计开发的基于动态问题及答案数据库的Answer Web自动答疑系统、北京师范大学vclass教学平台中的Askme！答疑部件等。目前我国实现自动答疑的方式主要有三种：第一种类型是预先在知识库中存储大量已关联好的问题和答案，通过关键词的匹配将答案直接返回给用户；第二种类型是将用户输入的问题语句构成集合，计算句子语义相似度，在已建立的问题语句集中找到相似的问题，并将该问题的答案呈现给用户；第三种类型是对用户的提问进行分词加权处理，提取出问题关键词，并通过全文检索匹配答案，然后返回给用户。

目前TREC[3]会议只提供面向英语的问答系统测评，但不可否认的是自动问答愈加成为人们关注的热点，ACM、SIGIR、ACL、TAC等国际会议都为问答系统提供了展示和交流的平台，推动着自动问答研究的不断进步。

四、论文的内容和组织结构

要设计一个面向指定科目的自动答疑系统，关键的技术包括关键词的提取、中文分词算法的选择、全文检索的实现以及知识库的构建。本文在研究了多种自动答疑系统后，分析目前答疑系统的研究现状及存在的问题，在Lucene系统架构的基础上，着重研究基于分词字典的匹配与频率统计的分词算法，综合两种算法的优点，在保证查全查准率的同时，高效的进行分词。

主要的研究内容如下：

1.阐述了自动答疑系统在网络教育中的重要作用，介绍了国内外典型的自动答疑系统的答疑机制和答疑部件，分析系统的当前研究现状和存在的问题。

2.介绍自动答疑系统所涉及的理论技术，比较三种全文检索技术的检索性能，提出并选择基于Lucene的全文检索引擎，对Lucene的系统结构、运行机制和系统自带的分词模块进行概述。

3.阐述了中文分词技术的发展现状和分词过程中遇到的技术难点，对常见的中文分词算法的优缺点进行分析，根据自动答疑系统中学生用户提问的特点，选择适合的分词算法并对分词性能进行改进。

4.分析知识的组织形式，合理构建知识库，对知识库中问题和答案的来源进行分析，提高查询和检索的效率。

5.对自动答疑系统的可行性进行分析，根据使用者的不同，满足不同用户的需求，介绍答疑系统的功能组成模块及总体设计模型，组织建立知识库，为系统的详细设计提供参考。

本文分为七章，各章的主要内容如下：

第一章引言，论述了本文的研究背景和研究意义，阐述了自动答疑系统的研究现状，介绍本文的组织结构。

第二章基于 Lucene 的全文检索技术，提出基于 Lucene 的全文检索引擎，对 Lucene 的系统结构、运行机制和自带的分词法进行概述。

第三章中文分词技术，论述了中文分词技术的发展现状和应用，研究常见的中文分词方法和分词过程中的难点。

第四章自动答疑系统的系统分析与总体设计，指出开发自动答疑系统的可行性，对答疑系统进行需求分析和总体设计，确立系统的设计目标和功能模块设计，完成系统数据库的设计。

第五章关键技术的解决方案，详细介绍了系统开发过程中几种关键技术所选用的方法，设计中文分词词典，组织建立知识库，实现 Lucene 全文检索技术和中文分词技术，消除歧义。

第六章自动答疑系统的实现与性能测试，介绍系统的开发环境和开发工具，根据设计方案完成系统基本功能模块的开发，测试了系统的准确率和检索效率，并给出了测试结果。

第七章结论，对本文研究的内容进行总结，对系统实际运行过程中出现的不足之处进行改进，对自动答疑系统的未来发展充满期待。

第二章 基于 Lucene 的全文检索技术

随着互联网的蓬勃发展和数据库技术的成熟，网络中存储着海量的信息并以几何级数迅猛增长，各种各样的信息无时无刻的不在充斥着人们的视线，只有对信息进行有效的整理才能真正为人所用，自动答疑系统的工作更加离不开对数据的频繁检索，为了准确、全面、快速的获取信息，必须使用良好的检索工具。比如之前百度发生的复大医院推广事件，宁波的一位患者想去复旦大学附属医院就医，通过百度查询之后搜索到的结果是“复大医院”，该患者选择前去就医后却发现自己百度到的“复大医院”并不是真正的复旦大学附属医院，百度表示这是由于两所医院的名称存在一定的语义相似性才出现的情况，并再度扩展了品牌保护关键词库。检索工具的选择对答疑系统的重要性不言而喻。

一、 检索工具的选择

信息检索的核心是全文检索技术，全文检索是一种允许用户使用自然语言根据数据文本进行检索的手段。全文检索技术由于查全率高、兼顾结构化与非结构化数据、检索速度快等优势在对信息的选择、分析过滤、安全管理等诸多领域提供了保障。通过对文献的研究发现国内对全文检索系统的研究大致分三类，分别是 MYSQL 数据库的全文检索功能、关系型数据库中 like 子句的模糊查询功能以及对全文检索软件的开发利用。综合考虑全文检索软件的灵活性、开放性、可扩展性，比较了 TPI、TRS、TRIP 等国内著名的全文检索软件，本文提出了基于 Lucene 的全文检索技术。Lucene 全文检索引擎、SQL 的 like 子句、MYSQL 数据库都能实现全文检索，但他们的查全率和检索性能有所不同，本文设计一个表名为 pdfbooks 的数据表，用以存储 PDF 中每页的文本信息，并使用 PDFBox 提取 pdf 文档中的文本信息并存储到数据库中，通过三种方式进行全文检索，比较检索耗时和查全率。

（一）MYSQL 数据库全文检索的实现

MYSQL 从 4.0 版本开始具有全文检索功能，MYSQL5.6 以下仅支持 MYISAM 表的全文检索，与模糊查询相比具有高效灵活的特点，而且仅支持英文检索，如果需要对中文关键字进行检索，首先要将中文转换成英文的书写格式，可通过 urlencode、MD5、区位码等形式对中文进行转换，再用英文分词方法进行分词操作，以达到建立全文索引的目的。但在实际操作中，需要根据中文词语的长度不断调整 ft_min_word_len 参数的默认值，系统默认的字段为 4，即使这样也常出现占用存储空间过大、无关联的检索结果过多的情况。用 SQL 语句“select * from title where Match（search）Against（关键词）”可以对数据库进行全文检索，允许对自然语言模式和布尔模式进行全文检索，当不指定具体模式时，默认采用自然语言方式，“select * from DB_Name where Match（search）Against（‘关

键词’ IN NATURAL LANGUAGU MODE）AS score”语句可以将返回结果按照从高到低的相关度自动进行排序。其中核心代码为 select * from pdfbooks where Match(pagetext) Against(" + keywords + ")"用来获取关键词字符串，GetConn(user，password) 与数据库连接，record = getResultSet(SQL)得到记录集。

```
catch ( SQLException e) {
e. printStackTrace( ) ;
}
Date end = new Date( ) ;
long time_index = end. getTime( ) — start. getTime( ) ;
System. out. println( time_index ) ;
System. out. println( " MQL 数据库全文检索耗时:  " ) ;
System. out. println( " MQL 数据库全文检索命中结果: " + count) ;
} catch ( IOException e ) {
e. printStackTrace( ) ;
}
```

（二）like 子句全文检索的实现

SQL SERVER 从 7.0 版本起新增了全文检索功能。与使用 create index 语句建立常规索引有所区别，全文检索存储在文件系统中，对相同数据库中的每个表依次建立全文索引，每次只能对一个表建立索引，并将新建立的多个全文索引表添加组织为全文索引目录，经过指定索引字段、创建填充索引等操作完成索引的建立。在关系型数据库中利用 SQL 结构化查询语言，使用语句“select * from DB_Name where 字段名 Like '% 关键词%'”对结构化数据进行检索，在模糊查询过程中对所有记录进行遍历，并对关键词进行匹配。其中核心代码为 " select * from pdfbooks where pagetext like % " + keywords + " % " 获取关键词字符串，GetConn(user，password)、record = getResultSet(SQL)依次用来连接数据库、获取记录集。

```
catch ( SQLException e ) {
e. printStackTrace( ) ;
}
Date end = new Date( ) ;
long time_index = end. getTime( ) — start. getTime( ) ;
System. out. println( time_index) ;
System. out. println( " Like 子句检索耗时: ( 毫秒) " ) ;
System. out. println( " Like 子句检索命中结果数: " + count) ;
```

```
} catch ( IOException e) {

e. printStackTrace( ) ;
}
```

（三）Lucene 全文检索的实现

基于 Lucene 的全文检索的实现主要包括两部分：索引的建立和检索索引。使用 Lucene 索引数据库需要完成以下几个步骤：首先创建一个数据库表，导入或者添加记录；其次通过 JDBC 访问数据库记录以建立数据流；数据记录作为文档，记录里的字段作为域添加到文档中；最后创建索引，并在循环中添加文档，直至数据表记录结束。在建立索引时遍历数据库中的记录集并为需要索引的字段建立索引，这样保证在检索时可直接在本地的索引中进行搜索，无需访问数据库。

```
public static void indexdb( ) throws SQLException {   //索引的建立
Date start = new Date( );
try {
File indexpath = new File( Dest_Index_Path);
IndexWriter writer = new IndexWriter( indexpath , new
StandardAnalyzer( ) , true);
GetConn( user, password);     //连接数据库
String SQL = " select * from pdfbooks";
record = getResultSet( SQL ); //获取记录集，遍历 record 记录集，为
需检索字段做索引
DbIndexBuilder( writer);
writer. optimize( );
writer. close( );
} catch ( IOException e) {
e. printStackTrace( );
}
Date end = new Date( );
long time_index = end. getTime( ) — start. getTime( );
System. out. println( "索引耗时: ( 秒 ) " );
}
```

```
public static void QueryIndex( ) {              //检索的实现
try {
Date start = new Date( );
IndexSearcher searcher = new IndexSearcher( Dest_Index_Path);
Term term = new Term( " pagetext", "关键词" );

Query query = new TermQuery( term );
Hits hits = searcher. search( query );
int count = 0;
for ( int i = 0; i < hits. length( ) ; i + + ) {
count + + ;
}
Date end = new Date( );
long time_index = end. getTime( ) — start. getTime( );
System. out. println( " Lucene 检索耗时: ( 毫秒 ) " );
System. out. print( time_index);
System. out. println( " Lucene 检索命中结果数: " + count);
} catch ( IOException e) {
e. printStackTrace( ) ;
}
```

（四）实验结果

实验在导入的 PDF 文档中随机选取 4 个关键词进行测试，得到如下图所示的耗时表和查全率表。

表 1 三种检索方式耗时表

keyword	耗时（毫秒）		
	Lucene	数据库 Like 子句	数据库全文检索
Trigger	148	397	76
Normalization	78	393	47
日志文件	78	443	48
数据挖掘	92	463	49
平均时间	99	424	55

表 2 三种检索方式查全率表

keyword	命中结果（个）			查全率	
	Lucene	数据库 Like 子句	数据库全文检索	Lucene	数据库全文检索
Trigger	8	9	9	89%	100%
Normalization	6	6	6	100%	100%
平均查全率				94.5%	100%
日志文件	107	125	0	85.6%	0
数据挖掘	114	136	0	83.8%	0
平均查全率				84.7%	0

通过实验发现在数据量较小时，使用模糊查询对结构化数据进行匹配具有灵活性高、查全率高的特点，但当数据量巨大时，对全部记录进行遍历以及与数据库管理系统的多次交互都会导致检索速度低下，由于 SQL 语句无法对非结构化数据进行检索，对数据表的增删改等后期维护操作很困难。Lucene 的检索速度略逊色于 MYSQL 数据库全文检索，但大大快于数据库 Like 子句的检索速度，平均耗时仅为 Like 子句的百分之二十，但是 MYSQL 数据库全文检索的问题是只能对英文进行高效检索，无法实现中文的检索，而 Lucene 对中文关键词的平均查全率高达 84%，是用空间换取时间的索引检索，特别是对于具有大量数据的数据库，Lucene 全文检索是很好的选择。因此，综合考虑三种检索方式的优劣，本文选择 Lucene 全文检索作为自动答疑系统的检索工具。

二、Lucene 简介

Lucene[4]是一个全文检索引擎工具包，给出了用来索引的成熟、免费、开放源代码并用 Java 编写的框架。软件开发人员不仅能使用它对特定的全文搜索应用程序进行建构，还可以将 Lucene 集成到各种系统软件中，为系统软件提供搜索功能。Lucene 提供了一组能够用于预处理、过滤、分析、索引和检索排序等操作的 API。作为一个全文搜索引擎，Lucene 具有许多优点：首先，8 位字节存储的索引格式可以独立于应用程序平台；其次，不依赖数据库的倒排索引[5]与分块索引的结合达到了优化提升索引速度的目的，小索引的建立与原索引的合并达到优化的效果；最后，Lucene 中默认实现了布尔查询、模糊查询、精确查询、分组查询等，而无需构建额外的信息检索模型。

三、Lucene 系统结构

Lucene 面向对象的系统架构使得扩展新功能变得方便容易，Lucene 的系统主要由三

部分组成：对外接口、索引核心、基础结构封装[6]，其中索引核心部分是重中之重，Lucene索引文件的结构类似于关系型数据库中表的行结构，由若干段构成，每个段又由包含多个域的文档组成，每个域中包含两个属性分别是域名字和域内容。

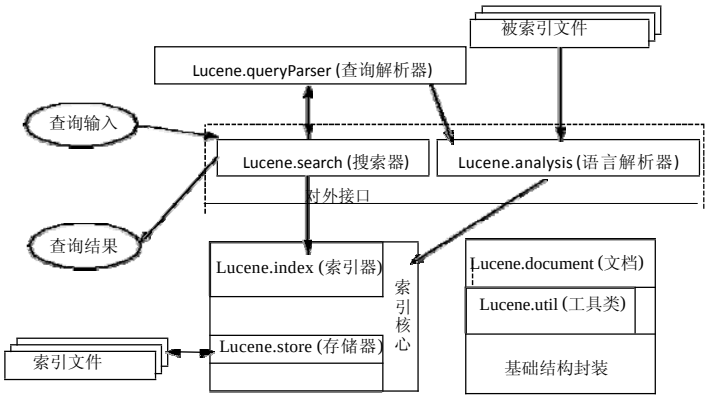

图 1 Lucene 系统结构

Lucene 将源码分成 7 个 package 子包，analysis 是用于各种语言切分词的语言解析器[7]，分词得到关键词和所在域值为组合的若干项，在 Analyzer 中项以 Token 的结构出现，它可以是单字、词组或短语，是索引中最小的信息单位，其中保存着关键词、关键词出现的频次以及关键词出现的位置等信息，在 Analyzer 中项就是分析源文件后返回给检索器的结果，本文设计改进的中文分词算法就是在 analysis 子包中完成的。lucene.search 的主要功能是索引管理，它通过收集用户的查询请求并根据查询条件搜索相关结果，提供索引接口来收集相关查询结果。lucene.index 的功能是索引管理，它为库提供了读写接口，库的创建、索引的建立、添加、更新、删除读取记录等操作都与该子包有关。Analysis、Search、Index 是 Lucene 中最重要的三个子包，是 Lucene 系统的核心。此外，还有一些子包与 Lucene 系统的操作密切相关，查询分析器 QueryParser 实现查询关键词间的运算操作；Store 为数据存储管理；Document 完成索引在存储过程中的文档结构管理；Util 为公用工具类。

四、Lucene 的运行机制

Lucene 通过构建索引库和检索索引库两部分实现全文检索。在构建索引库时，首先将要索引的文本内容通过分词处理切分成关键词，依次索引关键词，生成倒排索引结构的文件，完成索引入库；在信息检索中，先将用户输入的问题切分成词，生成以最小单位项为基础的查询对象，然后将查询条件传递给搜索器，并通过调用 lucene 查询解析器 queryParser 解析查询条件，访问索引库以查询符合条件的特定文档。基于 Lucene 中的向量空间的排序算法按照相似度从高到低进行排序，返回结果集 lucene.search.Hits 得到查询结果，Hits 提供了检索查询结果的缓冲，为结果的展示和返回提供支持。

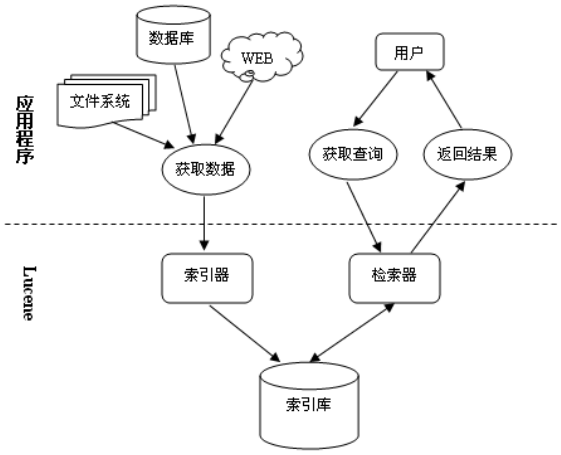

图 2 Lucene 全文检索过程

五、Lucene 的中文分词

Lucene 在具有强大检索功能的同时，在中分分词方面存在很大的缺陷。Lucene 中自带了多种分词工具，如 WhitespaceAnalyzer、SimpleAnalyzer、StopAnalyzer 等，但这几个分词器都是根据西方语言特点设计开发的。其中具备中文分词能力的主要有 StandardAnalyzer、IK_CAnalyzer、PaodingAnalyzer、MMAnalyzer、CJKAnalyzer，其中 CJKAnalyzer 是一个基于 Java 的中文日文韩文的语言分词工具包，它提供了 ChineseAnalyzer、CJKAnalyzer、SmartChineseAnalyzer 三种分词方法，但分词效果并不好，在对字符串“中华人民共和国”进行分词时，分词结果为“中/华/人/民/共/和/国”、“中华/华人/人民/民共/共和/和国”、“中华/人民/共/和/国”，这种分词没有意义还严重影响了系统后续的答疑工作，考虑到 Lucene 没有成熟的中分分词模块，并且自带分词器的准确性达不到答疑系统的要求，本文在 Lucene 系统架构的基础上，提出改进的中文分词算法，并将算法应用于 Lucene 系统中以提高系统的准确率。

第三章 中文分词技术

为了使答疑系统具有智能性，将用户输入的自然语言转化为计算机能理解、运用的语言，为了分析和处理用户提出的问题，我们必须首先研究中文分词技术[8] 。在中文语句中出现两个或两个以上的字，就可以组成一个词，同一个词在结合上下文的语境时经常表达出不同的意思，错误的分词方法会造成歧义，对答疑系统检索信息的正确率产生影响，出现返回给用户的答案与问题不一致的情况。

一、中文分词技术现状

早在 1960 年左右的俄汉翻译机发展时期[9]，苏联学者率先提出了“6-5-4-3-2-1”的分词方法，这也为中文分词技术的发展奠定了基础。近年来，随着 Internet 技术的发展，对中分分词算法的研究愈发成熟，海量信息技术有限公司从 1999 年开始从事中文语义的研究，采用复方分词法并与多种算法相结合综合处理问题，是目前中文语义识别领域最领先的企业之一；中文分词技术是百度、谷歌等搜索引擎的核心，搜索引擎的准确度与中文分词技术的准确性密切相关，百度自 2001 年开始自主研发了中文分词系统；许多高校和研究机构，如中国科学院、北京大学、上海交通大学也致力于研究中文分词技术，并取得了良好的反响。

二、常见的中文分词方法

目前，我国常见的中文分词方法有三类[10]分别是基于字符串匹配的分词、基于统计的分词以及基于理解的分词。

（一）基于字符串匹配的分词法

基于字符串匹配的分词法[11]离不开分词词典的帮助，由于该种分词法具有算法简单、易于理解、容易操作上手、分词速度快、使用范围广等诸多优点，它是分词算法中出现时间最早，发展最为成熟的算法，也是当前使用频率最高的方法。根据特定的分词机制，将待切分的语句与包含庞大数据的分词词典进行匹配，若匹配到字符串则认定成功识别出一个词，按照这种方式将待切分的汉字串与分词词典进行频繁匹配，直至切分出全部的词，在词与词间插入分隔符作标记并将处理过的字符串输出完成分词操作。但由于单一分词法的分词精准度无法满足实际答疑系统的需要，还可以将基于词典的分词法与其他方法相结合，达到减少切分歧义的目的。基于字符串匹配的分词法对词典依赖性非常高，分词词典的选择以及词典中词条的数量和词条的覆盖范围是否全面将直接导致分词的成功或失败，但目前没有权威的词典对词进行规范，容易造成切分歧义，无法对新词和未登录词进行有

效识别，且分词过程中需要与词典频繁进行匹配，严重影响了分词速度。

（二）基于统计的分词法

基于统计的分词法在统计思想的帮助下，分词可以在不需要借助额外的分词词典的情况下实现，因此它通常被称作无词典分词或统计取词法。从汉字组词的特点来看，汉字中两字词语最多占总词语的百分之六十以上，其次数量较多的为三字词、四字词等，因此可以利用统计算法对提问文本中相邻单字的组合情况进行计算，计算出的结果越高意味着相邻两个字组合成一个词语的几率越大，这种对待处理文本中相邻字的出现频率进行统计的方法称作互现信息。互现信息用来表示汉字之间关系的紧密程度，当紧密程度大于设定的阈值时，确定相邻的两个单字成为稳定的组合，可以构成一个词。[12]该方法的缺点是系统资源开销昂贵，还存在着高共现频率的组合不是词的情况。在实际分词过程中常将基于词典的匹配与基于统计的分词法相结合，综合二者的优点的同时对交集型歧义有良好的识别和消歧的能力。而基于统计机器学习的方法，需要建立相应的数学模型，将按照汉语组词规律分好词的文本集合作为统计模型建立的基础，语料库的规模、准确性、代表性与分词效果息息相关，利用统计、模式匹配和机器学习等方法分析规律、训练参数，建立高效的分词统计模型。

（三）基于理解的分词法

基于理解的分词法是使计算机尽可能模仿人的思维特点，从汉语断句的习惯、上下文语境的结合以及人名、地名、机构名等固定词语的积累方面着手，对待处理的语句在分析理解的基础上进行切分，主要对基本分词操作中出现的歧义结果进行处理，选择符合人类思维和理解特征的结果作为正确的切分结果，达到明确分词的效果。[13]这种分词法由三部分组成：分词子系统、语法和语义子系统以及通用控制部分。在通用部分的控制协调下，分词子系统可以获得与文档有关的词语和句子的语义信息和句法信息，句法和语义子系统通过对人分析理解过程的模仿判断出正确的分词结果。它主要包括联想回溯方法、联系上下文语境法、专家系统等分词方法。由于个体的差异，不同的人对语法句意的理解也不同，这对建立知识规则增加了难度。基于理解的分词法目前处于实验研究阶段的初级阶段，距离实际应用还有一段距离，但不可否认的是基于理解的分词法对中文分词法的研究进步具有极大的促进作用。

三、中文分词的难点

汉语是一种非常复杂的语言，在形声义等多方面蕴含着不同的信息，尤其在特定的词句、语境中更表达出不同的情感，汉字是中华民族的象征更是古人流传下来的瑰宝，具有

丰富的历史文化底蕴这些都毋庸置疑，正是由于这些原因对中文语句的切分造成了极大的难度，在分词过程中一直存在着两个问题困扰着研究者们，即歧义的识别消除和新词的识别。

（一）歧义识别

歧义意味着对相同的句子进行断句分割可能存在两个或更多的分词结果，选择的分词策略与切分算法不同，得到的切分字句也不尽相同，但根据人的分析习惯，结合上下文的语境描述，只有一个正确的切分结果符合当前要求。根据文献统计，在文本分词的过程中出现模糊的概率约为 1/110[14]。常见的分词歧义包括交叉歧义、组合歧义和真实歧义。交叉歧义指的是相邻两词之间有重叠的部分，如“阿朱原来生活在这里一段时间”中“原来”和“来生”共用了一个“来”字，“来生”和“生活”重用了一个“生”字，在歧义识别的过程中交叉歧义出现的次数很多，可以使用统计的手段进行消除。与交叉歧义相比，组合歧义的发现与消除更让人伤脑筋，组合型歧义意味着词语的一部分也是一个完整的词，而且这种歧义是基于整个句子来判断的。如“中共中央委员会”、“中国人民代表大会”中“中共”、“中央”、“委员会”、“中国”、“人民”、“代表”、“大会”这些都是词，但当它们组合在一起时也构成一个词，考虑到在分词算法中常选取最大词长对句子进行匹配，所以正确的分词结果应该将“中共中央委员会”、“中国人民代表大会”作为完整的词切分出来。真歧义是指给出一个句子，如果不联系上下文，人们也无法判断出哪些是词，哪些不是词，如“新生市场经过长时间的调整终于进入了旺季”中“新生市场”一词如果不联系前文很难知道其真实的含义究竟是“新入学的学生市场”还是“新产生的市场”。另外，一词多义使得相同的词在联系不同的上下文中产生的分词结果也不同，这使得对计算机的区分识别变得更加困难。

（二）新词识别

新词识别特指未在分词词典中出现的词[15]也称作未登录词，包括人名、地名、机构名、新生词、货币名、衍生词、缩写、省略语、专有名词等，没有任何一部词典可以包含全部的词条，然而未登录词的识别对分词结果的有效性具有巨大影响。目前未登录词可以通过三种方法找到，分别是基于规则的识别、基于统计的识别、基于规则与统计结合的识别。基于规则的识别总结汉语的构词规律，制定未登录词的识别规则，这需要预先建立好识别规则集，而且由于汉语词汇的不断衍生和频繁变化，规则集的可移植性较差，不利于系统的发展；基于统计的识别方法利用概率论的知识，运用预先建立的语料库对统计模型进行训练达到理想的效果，通对过相邻中文字符串的互信息和中文字符串的特定特征信息进行统计，可以判断汉字字符串是否是未登录的字，统计识别的方法提高了可移植性，但没有考虑到语言学知识和汉语构词的规律，在准确率上有所欠缺；考虑到两种识别方法各有千

秋却可以相互补充，可以通过二者结合以充分发挥各自的识别优势。

四、中文分词技术的应用

中文分词技术作为自然语言理解的领域和信息处理的基础，不仅大大应用于搜索引擎中，也不断拓展于机器翻译、自动分类、简繁体转换、分类校对、自动摘要等多方面。Word还提供了自动标记文本、错词校对、对文字标音的功能；计算机的系统多与接口相关，汉语的接口遵照人机交互的理论更要依赖中文理解才能建立，国外的计算机处理技术若想打入中国市场先要解决中文分词的问题，这给中国的企业机构带来了商机。中文分词技术的应用也不仅仅局限于文本输入的模式，生活中无论是微信聊天中的语音与文字间的转换，还是语音导航时输入的语句，甚至是智能家居的普及都与中文语义的理解有着莫大的关系，众多的电子产品要想体现出智能性，也离不开中文分词技术。此外，对语言文字的处理，包括句型识别、词频统计、词结构分析等问题也有帮助，中文语义的理解与我们的生活密不可分，渗透在日常生活的方方面面，只有不断的提高中文分词技术的性能才能更好的应用于社会服务于大众。

第四章 自动答疑系统的系统分析与总体设计

一、系统可行性分析

作为课堂教学的补充和延续，自动答疑系统为教师和学生提供了一个没有时间和空间限制的交流场所，答疑系统可以不分昼夜、不知疲倦的为人们服务，学生在学习过程中出现的问题可以及时被解答，这将大大的减轻了教师重复、繁重、无意义的答疑工作。面向特定科目的答疑系统，由于课程的内容不会频繁的更改，可以根据教学大纲的要求对少数内容的变更进行定期更新和维护，灵活的进行管理。学生通过答疑系统解决难题和困惑提高学习效率，与同学之间的交流促进了知识的理解，教师辅助答疑进一步把握学生的学习进度和对重难点的理解掌握。针对师生的共同需求，本文设计的面向特定科目的自动答疑系统是可行的。

二、需求分析

自动答疑系统的用户是教师、学生、管理员，为满足用户不同的需求，根据角色的不同赋予其特定的操作权限，设计不同的功能模块并实现功能。

从学生的角度看，主要使用该系统回答难题。用户注册登录后，在显示的搜索框中输入自然语言进行提问，若知识库中有相应的问题，系统将简洁正确的答案返回给学生；若知识库中无类似问题，学生可以进行提问，系统将未解决的问题统一转交由教师进行回答。学生可以通过答疑论坛的问题列表查看其它同学的疑难问题，并进行回答，学生回答的问题不会被系统直接采纳，但可以给其它同学提供新的思路，参考了其它同学的答案后，可以对答案是否有所帮助进行评价，在一问一答的交互过程中，既解答了学生的实际问题又增添了学习兴趣与探索的好奇心，同时促进了同学之间的交流，提高了使用者的主观能动性和自主性。

在教师看来，老师在给学生传授知识的教学过程中，主要引导学生培养获取知识的途径以及如何更好的接受所学知识。教师回答系统中学生提交的未解决问题，提交后系统自动将此问题答案添加到知识库中，并将已解决的问题通过系统消息提示学生进行查看。教师通过回答疑难问题并对相关问题进行总结，了解学生对课程的理解程度和学习进度，方便进一步安排教学内容与教学过程，更好的完成教学目标。

在系统中承担管理用户、词典、知识库任务的是管理员，他拥有最高权限。管理员需要对系统数据库进行定期备份，防止数据丢失，对用户的信息进行管理，对知识库有增加、删除、修改、维护的工作。管理员的系统管理包括版块列表、管理员列表、用户列表、系统公告列表、问答列表、问答首页等模块，其中待添加的问题可以按版块列表中的类型进行分类，问答列表中存储着来自系统问答和提问采纳两种来源的问题—答案信息。

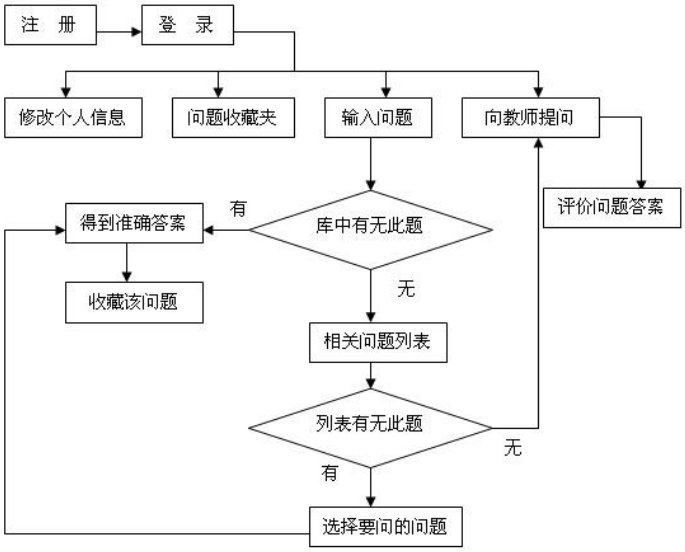

图 3 学生业务流程

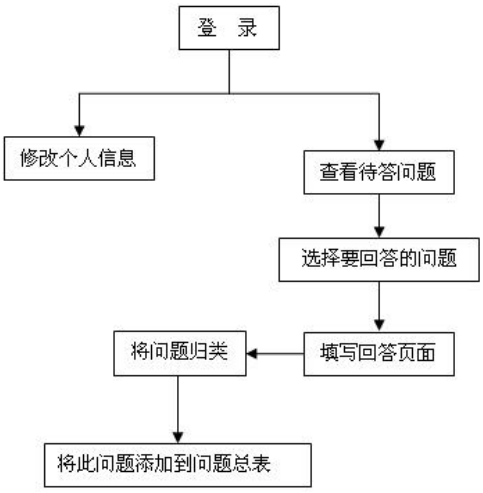

图 4 教师业务流程

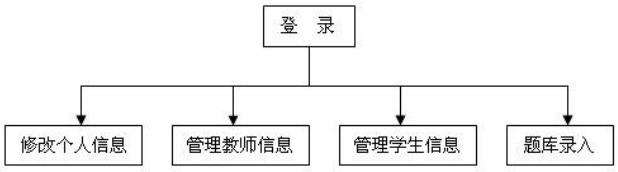

图 5 管理员业务流程

三、系统的设计目标

本文的研究对象是基于网络面向指定科目的自动答疑系统，实现其中中文分词模块、答疑模块、用户界面、系统架构等关键技术，在中文分词技术、检索方式、系统的查全查准率方面进行改进。该系统支持使用者同时采用关键词和自然语言两种形式来提问，并获得实时准确的答案，运行良好的自动应答系统应具有以下特征：

1. 系统的人机对话窗口根据使用者的不同提供具有不同功能模块的操作界面及不同范围的使用权限。

2. 系统可以对用户自然语言的提问分析理解，通过搜索知识库，返回给使用者按照从高到低的相关度进行排序的正确简洁的答案。

3. 系统提供针对特定科目的问题和答案，提供给使用者如系统答疑、辅助答疑、在线论坛等多种形式。

4. 确保系统具有良好的稳定性、持续性和移植性，同时加强对使用者的身份审核和严格的权限控制。

5. 定期更新和维护系统知识库。

四、功能模块设计

（一）知识库的组织

知识库中的知识通常由具有相互依赖关系的模块化的知识片构成，它在人机交互的过程中与数据库相结合，通常采用分层存储结构，最基础的是基于底层事实知识；中间过渡层是用来控制事实的知识的规则和过程，其中规则是最常用的典型知识；最高层是决定控制中间层知识，这被认为是规则的规则。常见的知识表示方法包括问句归约法、谓词逻辑方法、状态空间方法、语义网络方法等[17]。

知识库中的问题答案信息和使用者的信息数据共同构成了自动答疑系统中的数据。其中知识库信息数据的来源是在系统建立之初，由领域专家和教师根据教学大纲和实际教学经验从书本、网络等多种途径整理出该领域的常见问题和答案，构建成为系统初步的知识库；另一来源是在系统运行过程中，教师对学生疑难问题的解答对知识库的扩充，以及日常维护操作，如根据学生提出的问题，教师对系统添加删除和更改知识库等。

（二）答疑系统运行方式

如今，常见的自动答疑系统主要分为根据关键词查询和通过自然语言提问两种形式。仅使用关键词进行匹配搜索的系统，可以对数据库中存储的结构化数据进行匹配，但对存储在磁盘上的 pdf、doc、docx、xls 等非结构化数据无计可施。同时，采用关键词查询的系统一般不支持长语句和自然语言的提问，无法准确延伸出用户实际的提问意图。

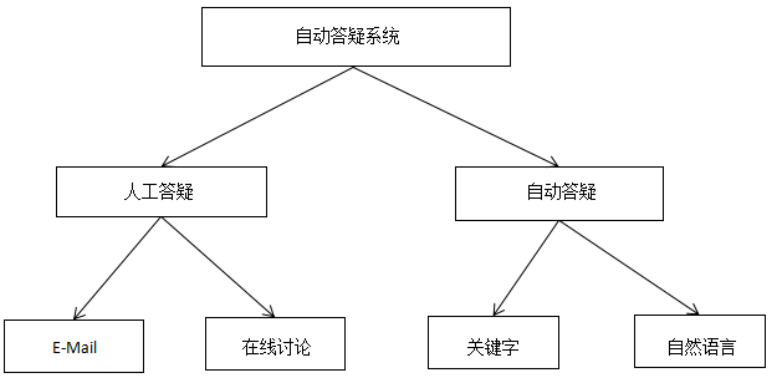

图 6 自动答疑系统的答疑方式

当系统自动回答问题时，系统首先分析和预处理使用者提交的问题陈述语句，若用户输入的内容是关键词，则直接在知识库中进行检索；若用户采用自然语言提问，系统先对问题语句进行分词处理，提取问题语句的关键词，再通过 Lucene 检索得到答案并返回用户。若自动答疑的结果使用者不满意，可以选择人工答疑，用户可以在系统中发布未解决问题，系统自动向领域专家和教师提问，相应的问题呈现在教师待解决问题的列表中，教师登录系统对按照时间顺序排列的待解决问题进行回答，教师回答完毕后系统发送消息通

知学生，学生点击查看已解决的问题和答案。与此同时，系统也将未解决的问题呈现在答疑论坛中，其他学生可以对未解决问题进行浏览和回答，提出自己的想法，学生们在答疑论坛中进行在线讨论，从讨论中得到新思路，选取满意的答案。

（三）功能模块设计

根据系统的需求分析和设计目标，系统应具有注册登录、自动答疑、信息管理、系统公告、答疑论坛等模块。

1. 身份验证模块。系统中的相关信息只对被授权的用户开放，身份验证模块的作用是检验用户账号的合法性，只有通过身份验证的合法用户才能访问系统。

2. 自动答疑模块。分词是对用户输入的自然语言问题的分词，提取出关键词，基于所获得的关键词在知识库中进行检索，借助 Lucene 全文搜索技术返回满足用户要求的答案。

3. 在线答疑模块。在线答疑模块由系统公告、答疑论坛和邮件答疑组成，系统公告中显示管理员发布的各项信息，包括系统通知信息以及提示学生查看已解决问题的消息；如果检索返回的结果无法使用户满意，可以将问题发布在答疑论坛上展开讨论，从中选取合适的答案，用户还可以对答案是否产生了帮助进行评价；在自动答疑和答疑论坛都无法得到答案的情况下，可发送邮件直接向教师提问。

4.信息管理模块。信息管理模块由管理员和教师定期维护，包括用户身份审核、使用者信息更改、知识库的添加和删除、在线论坛的维护以及消息的管理。

五、系统数据库设计

（一） 数据库的概念设计 E-R 图

1.问题录入子系统的 E-R 图

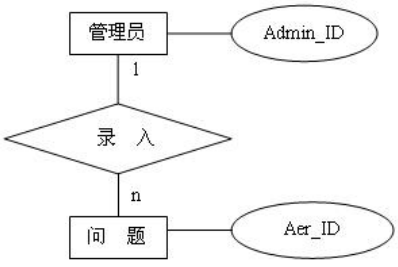

图 7 问题录入子系统的 E-R 图

2.问题检索子系统的 E-R 图

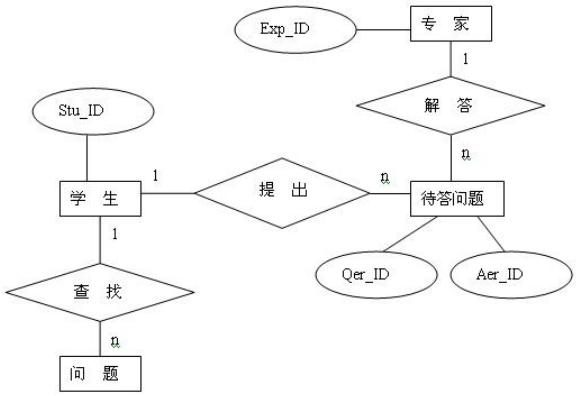

图 8 问题检索子系统的 E-R 图

3.系统总 E-R 图

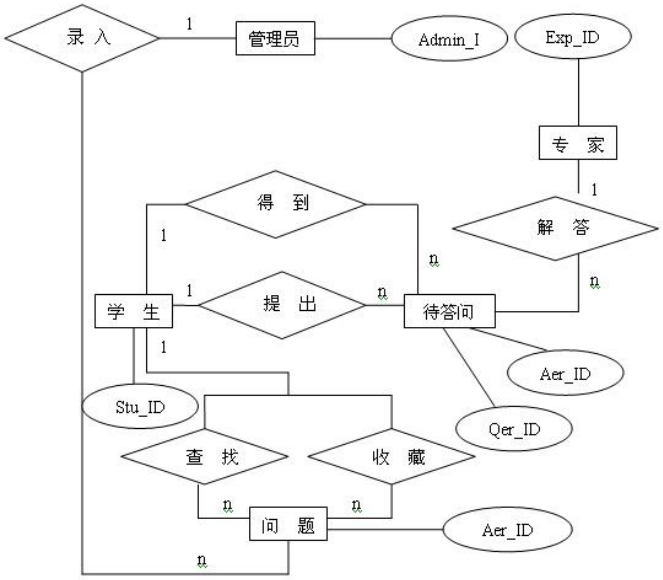

图 9 系统总 E-R 图

（二） 数据库的物理设计

数据库中的关系模式由 Admin、领域专家或教师、Student、问题-答案、教师待回答问题、问题-答案的类型编码等多个方面组成，根据关系模式在数据库中创建相应的数据表。

Admin 表用来记录管理员 ID、用户名、密码等基本信息；领域教师 Teacher 表中存放着教师 ID、用户名、密码在内的专家基本信息；Student 表记录着学生的基本信息，包括学生的学号、用户名、密码、待回答问题、已解决问题、问题收藏夹、邮箱等信息；QIdentity 表和 AIdentity 表中分别存储的是问题类型编码、答案类型编码，数据表中 Qtype 字段和

Atype 字段的功能分别是对问题类型的标识进行编号以及对答案类型的标识进行编号，Identity 字段中保存着答案字段的名称；Question-Answer 表中记录着问题 ID、答案 ID、回答问题的领域专家教师的 ID、问题-答案的所属类型、对答案是否有帮助的评价等与问题相关的信息，其中问题-答案的所属类型按不同的类型相区分，不同数字代表不同的问题-答案的所属类型，其中 0 表示提问的类型属于概念问题，1 表示提问类型为功能类问题，2 表示提出的问题为比较事务间的区别，3 表示提问类型为实现的方法；Wait 表对待回答问题的信息进行存储，主要包括待回答问题的 ID、提问者的 ID 和用户名、回答问题教师的 ID、提问者的提问时间、领域专家教师的回答时间、问题的答案等。在数据库中生成上述的关系数据表，得到如下图所示的数据库关系图。

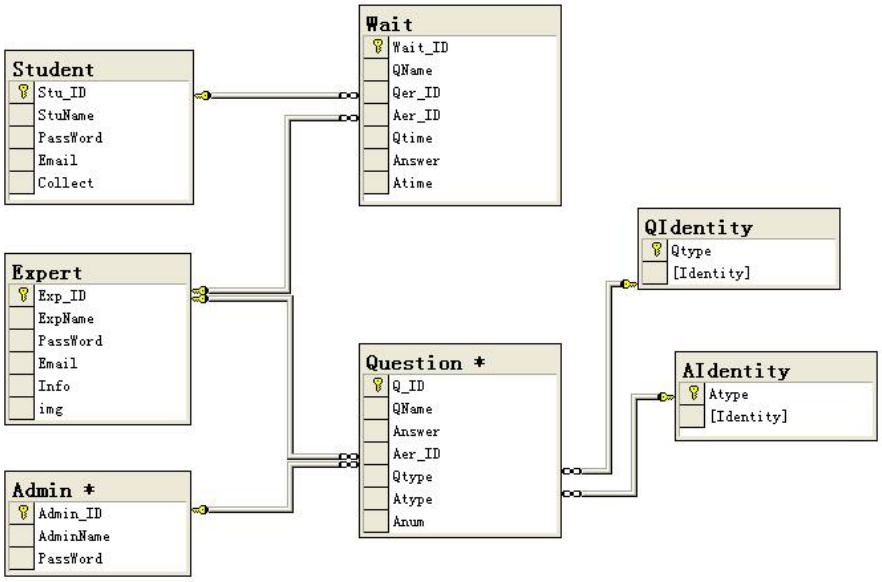

图 10 数据库关系图

第五章 关键技术的解决方案

一、Lucene 全文检索的实现

基于 Lucene 的全文检索引擎具有开源、查询能力强、检索速度快的特点，可以在此基础上进行二次开发。本文开发一个基于网络的在线自动答疑系统，该系统采用 B/S 结构为学生提供了快速交互应答的场所，而且不受时间和空间以及周边环境的局限。与数据库中对记录建立索引，模糊查询时需要遍历所有记录的检索方式不同，Lucene 中完善的倒排索引全文检索技术是对关键词进行索引的检索形式，采用字或者词作为索引项，它保证了答疑系统的查全率，它类似于通过浏览字典中的拼音字母表或部首笔画表查询定位汉字的过程。一般的正向索引建立的是从文档到关键词的映射方式，而倒排索引建立从关键词到文档的映射，文档按所在文件库的编码、位置和偏移量分别表示，在索引结构中分别存储关键词、关键词出现的次数以及关键词出现的位置三种文件信息。在记录了每个词出现的频率后，采用倒排索引在文档结构中依次索引词语，方便系统的检索。[18]此外，Lucene 提供的是搜索内核，它不关心数据格式、来源，无论是否为结构化数据，只要它可以转换成文本格式，就可以进行检索，Lucene 能够处理的内容包括互联网上的 Web 页面、Office 文档、HTML、PDF 文件等。

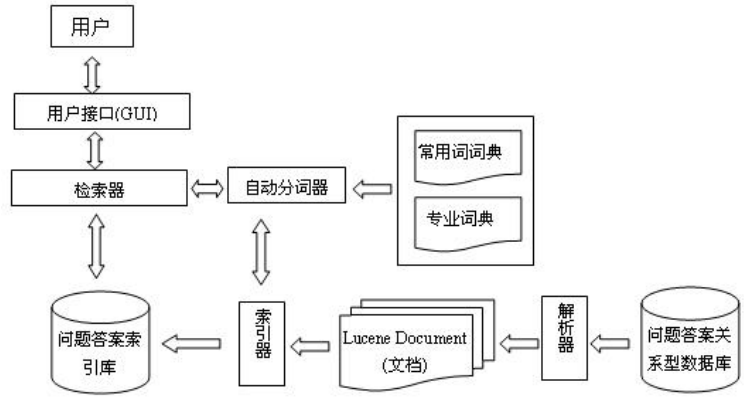

图 11 基于 Lucene 答疑模型的建立

（一） 索引的建立

要实现 Lucene 全文搜索功能的首要工作是建立索引，索引的作用是对搜索器要检索的信息进行分析处理，抽取出一连串用以表示文本信息的索引项并生成索引表。考虑到某些词组、字串在问题中频率出现，但它们并不表达实际意义，例如：“什么、的、为什么、如何、了、是、吗、怎么”等等，首先对文档进行预处理，去掉标点符号和无意义的停用词。程序在执行过程中依次调用 Preprocess()、charactor Process()、split To SmallFiles（）三个函数进行字符处理和文档切分，完成对文档的预先处理工作。在索引时，Lucene

使用 IndexWriter、Document、Field 这三个类，IndexWriter 对数据进行传递和分析使其更适合被索引，Document 指定索引内容并结合分词算法进行分词，构建索引库时，首先获取问题答案库中的数据，将其转换为 Lucene Document 结构并生成反向索引结构的索引文件，以完成索引存储。

建立标准的 Lucene 索引的方法为 public void index(String sqlTest, boolean isConsistent) throws Exception，其中关键的语句为 IndexWriter writer = new IndexWriter(ramDir,indexanalyzer, isConsistent)其中 ramDir 是索引库的位置，isConsistent 确定是否初始化索引库， false 为已存在库建立增量索引。Lucene 通过建立分组索引文件的方法，保证用户能及时检索到最新答案，当分组文件达到一定数量时，调用 writer（）函数将全部的分组索引文件合并为一个，这样操作的好处是减少了频繁的 I/O 操作。

（二）检索索引的实现

索引的建立是为检索提供服务，Lucene 的检索主要利用 3 个类实现，它们分别是 QueryParser 查询解析器、索引搜索器 IndexSearcher 和 Hits ，使用者输入的提问语句需要经过 analysis 语言解析器中的分词算法，切分成索引中最小的存储单位项时，才能作为查询对象进行信息检索的操作，查询条件成功传入搜索器后，需要经过 lucene.queryParser 查询解析器对查询条件进行调用解析，访问索引库从中得到到符合条件的查询结果，返回结果集 lucene.search.Hits 得到一组按相似度排序的查询结果。对关键词进行高亮设置的工具包是 org.apache.lucene.search.highlight，使用 new SimpleHTMLFormatter("", "")语句将知识库中与学生提问语句中相同的关键词进行高亮设置，使检索结果更加醒目。Lucene 中自带的基于向量空间模型的 TF/IDF 排序算法与评分机制，对检索结果按照从高到低的分数来排序，与用户需求相关度较高的答案排在前面并返回给用户，其中 TF 为关键字频率，IDF 为逆文档频率，文档的加权值为 Boost、文档关键字与文档的长度比为 LengthNorm。在计算相似度时，主要通过调用 Similarity.idf(Term,Searcher)计算文档的 IDF 值,通过 Query.getBoots( )得到权重，在 Scorer.score( )中计算得到的数据包括了检索关键词的频率、关键词的文档出现频率，以此得到最终的相似度得分。

二、设计分词词典

由于基于统计的分词法和基于理解的分词法在建立相应的数学模型的同时，还需要大量的文本信息作为分析法则、训练参数的基础，为了实现高效的分词，这两种分词方法难易操作，不易达到理想的分词效果，因此目前使用最多的分词方法仍为基于字符串的分词法。基于字符串的分词法依赖词典，汉语词典中大概有十几万条词条，如果每次分词操作

都与分词词典进行频繁比对，那么查询效率会十分低下，为了保证中文分词具有良好的分词能力，首要的工作是对分词词典进行设计，这包括选择分词词典的类型、内容、模型及词典机制。本文从减少待切分的文本数目和分词词典的匹配次数开始，提高了分词词典的匹配效率。[20]

（一）基于整词二分法的词典机制

整词和逐字二分法的词典机制相同。均是由如下图所示的首字散列表、词索引表和词典正文组成。首字散列表包括首字、入口项个数和指向词索引表的指针。词典中汉字的排序由首字的国标码确定，并与首字散列表中的序号相对应；词典正文是以词为单位的有序表，按照首字相同字串个数增加的形式进行排列。词索引表是指向词典正文中每个词的指针集合。通过首字散列表的哈希定位和词索引表可以确定词在词典正文中可能的位置范围，进而通过整词二分进行定位。

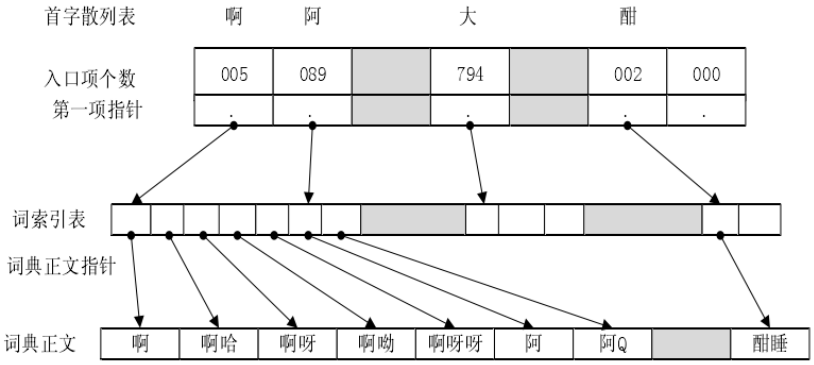

图 12 基于整词二分法的分词词典机制

（二）基于 TRIE 索引树的词典机制

TRIE 索引形式表示的键数，可以有效的建立数据检索组织结构，方便字符串的存储并实现快速查找。TRIE 索引树节点是按照关键词排序的数组，由三部分组成：关键字、子树大小和子树指针，即使对等待查询的字串长度一无所知，只要沿着树链逐字逐句的搜索到结果，最坏的情况是 O（n）其中 n 是索引树的层数，最快的情况是在第一层就检索到了关键字，时间复杂度为 O（1），其中比较明显的缺点是索引树的建立和维护消耗了大量的空间。

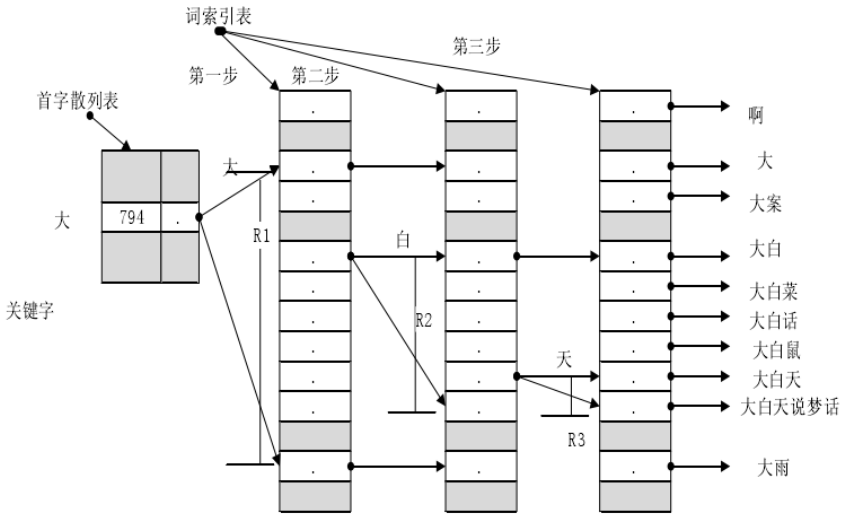

图 13 基于 TRIE 索引树的分词词典机制

（三）基于逐字二分法的词典机制

基于逐字二分法是在整词二分法和 Trie 索引树词典机制的基础上进行的改进，词典结构类似于前两个词典机制，在整词二分法中保留了首字散列表的第一个字、入口项个数和指向词索引表的指针的词典结构基础上，它还在查询速度方面利用了 TRIE 索引树的优势，并使用逐字匹配的方法来提高匹配的效率。

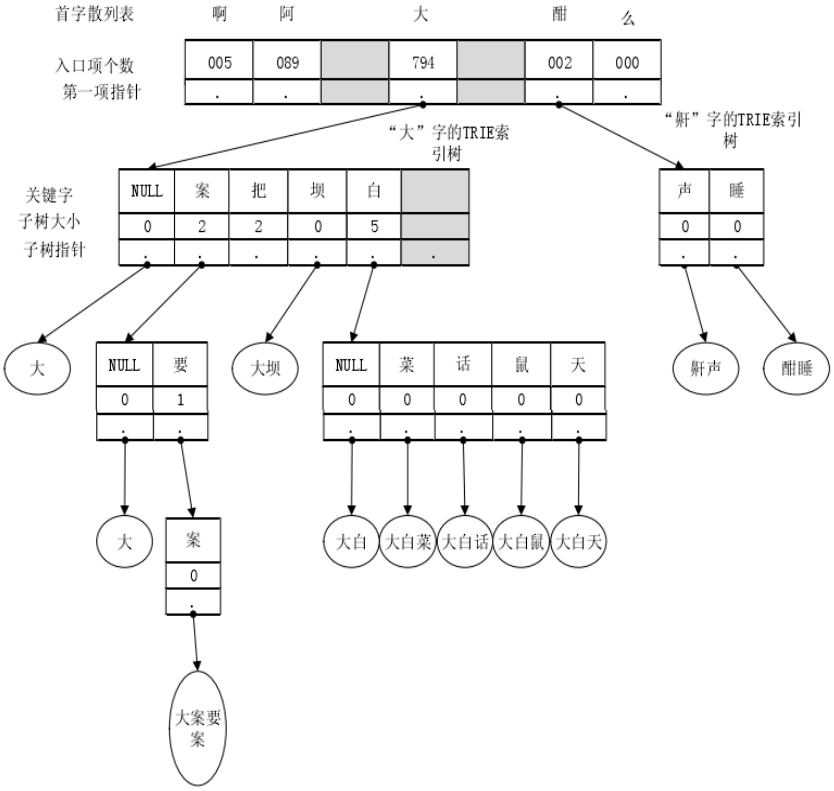

图 14 基于逐字二分法的分词词典机制

虽然这种词典机制被称作逐字二分，但它并不属于真正意义上的逐字二分的词典机制是因为没有对数据结构进行改进。在上述三种词典机制的基础上，为了提高分词速度本文采用双字哈希索引的词典机制，哈希词典的检索速度快，空间利用率高。

（四）基于双字哈希索引的词典机制

基于双字哈希索引的词典机制依次索引待索引字段的前两个字，形成深度为 2 的 TRIE 子树[23]。

首字 Hash 索引中的关键字存储首字字符，并使用 2 个字节作为存储单位；该标识位用 1/0 判断首字是否为词语，当首字为词语时，用计数器记录该词频率。次字哈希索引用关键字来存储次字字符，以 2 个字节作为存储单位；该标识位用 1/0 标识，用于判断第一个单字和第二个单字是否能组合为词语。剩余字串指针指向以首次字起始的所有词语剩余字串有序数组。剩余字串索引表包括剩余字串内容和标识位两部分，其中剩余字串内容以 2*n 字节进行存储，为删去字串的前两个字符的剩余部分，该标识用于标记剩余单字是否能构建成词[24]。这就使得两个字以下的词语可以使用 TRIE 索引树提高中文分词检索的速度，它对常规的词典机制具有检索匹配快的优势，三个字以上的词语通过线性表检索匹配，减少索引匹配的时间，提高分词的效率。

（五）分词词典设计

本文对分词词典的设计分为三个方面，分别为专业词典、常用词典和停用词典。对面向指定科目的答疑系统来说，最重要的词典是专业词典，专业词典中存放着面向特定学科的专业词汇，由于本自动答疑系统是面向数据库组成原理的答疑，涉及内容的专业性很强，则需要在专业词典中添加数据库学科的相关词汇，如“谓词”、“SQL”、“关系模型数据库”等。在实际的答疑过程中，用户不一定单单用中文进行输入，因此要考虑到用户输入的问题中是否含有英文、数字、符号，并对其加以区分。建立停用词典，系统对停用词典中出现的字词不进行检索，提高检索的效率。

在前文对常用的分词算法进行概述的过程中了解到，基于词典的分词法在具备分词速度快，算法简单、开销较小等诸多优势的同时为越来越多的研究者所喜爱，但容易出现歧义，难以解决未登录词的问题，而基于统计的分词法可以很好的解决歧义的问题。在答疑系统日常运行的过程中，综合选用基于词典的切词法与频率统计的方法，基于分词词典的切词方法，在系统空闲时间里，采集统计信息来维护词典，根据分词算法设定的频繁域值，当字符串序列超过此频繁域值时，认定该字符串为一个词。

三、中文分词的实现

目前，研究者常使用两种手段来提高中文分词的准确性和切分速度，这两种方法旨在改进分词词典和分词算法，降低要切分的提问语句与分词词典的匹配次数来提高分词速度进而完成对分词词典的优化和改进分词算法，以提高发现歧义减少歧义的能力以及对新词未登录词的识别能力。考虑到 Lucene 中没有成熟的中分分词模块且自带分词器的分词精度不达标，本文综合两种改进方式，为了找到可以与 Lucene 系统架构进行匹配，符合答疑系统准确率要求的中文分词算法。

（一）正向最大匹配算法

基于长词优先原则的分词法首先需要分词字典作为依据，对待切分的字串，从初始位置开始取最大词长为 m 的字串，如果该字串成功匹配分词字典中的词条，则该字串被认定成功识别一个词，如果匹配分词词典不成功，从字串的尾部去掉一个字重新匹配，直至匹配成功或剩余一个单字为止。一次分词操作结束后，将成功切分成词的下一个汉字作为起始，继续重复上述过程。若待切分的字串长度小于最大词长，则将字串长度赋值给 m,若仍匹配失败可以对最大词长 m 重新取值匹配。如果字符串无法与分词词典匹配，则将其作为未登录词进行处理。最大匹配算法的核心思想是将字符串与最大词长的字串进行比对，以便最小化词汇总量。

算法的核心代码如下：

```
public class FMMsegment {
public String[ ] segment(String str, int len,int max Len) {
int MAXLEN = max Len; // 最大词长
boolean flag = false; // 是否在词典中匹配成功
int n Word = 0; // 切分词的个数
String[ ] words = new String[len];
int begNum = 0;
int endNum = 0;
while (begNum < str.length( )) {
endNum = begNum + MAXLEN - 1;
String tempWord = new String( );// 临时子串
while (endNum > begNum) {
tempWord = str.substring(begNum, endNum);
```

```
flag = isInDictionary(temp Word); //调用 isInDictionary 方法判断词典中是否有这词
if (flag) {
words[nWord] = tempWord;
nWord++;
begNum = endNum;
} else {
endNum = endNum - 1;
}
begNum = endNum + 1;// 从此词的下一个位置开始
}
return words;
}
```

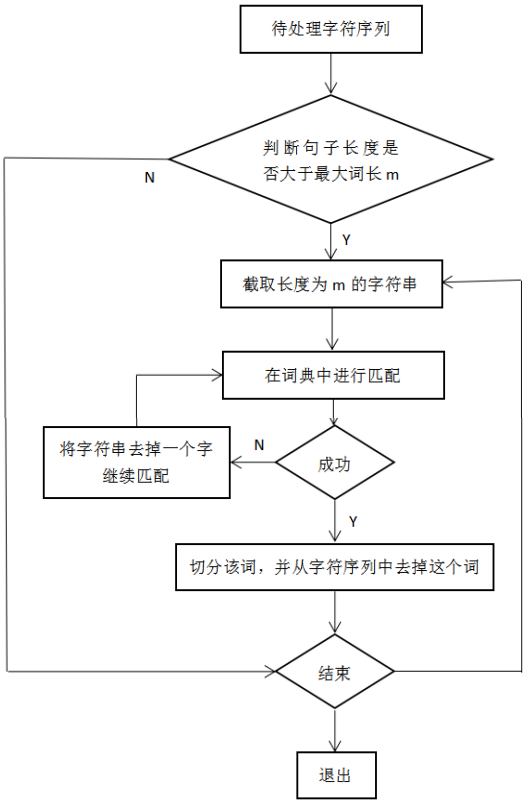

图 15 长词优先匹配算法流程图

如果采用正向最大匹配对“我们是中华人民共和国公民”这个文本进行分词，设最大词长为 7，则分词过程如下：

表 3 正向最大匹配算法的分词过程

步骤	待切分文本	操作	分词结果
1	我们是中华人民共和国公民	取七个字符	空
2	我们是中华人民	无匹配结果 字符串去掉一字	空
3	我们是中华人	无匹配结果 字符串去掉一字	空
4	我们是中华	无匹配结果 字符串去掉一字	空
5	我们是中	无匹配结果 字符串去掉一字	空
6	我们是	无匹配结果 字符串去掉一字	空
7	我们	无匹配结果 字符串去掉一字	我们
...	...	...	...
	是	与词典匹配成功， 切分出该词	我们\是
	中华人民共和国	与词典匹配成功， 切分出该词	我们\是\中华人民共和国
	公民	与词典匹配成功， 切分出该词	我们\是\中华人民共和国\公民

（二）逆向最大匹配算法

类似于前面所提到的正向最大匹配算法的分词步骤，但是文本以相反的方式进行扫描，采用逆向最大匹配算法开始扫描字串的末尾并作为初始位置，首先切分靠后的末尾部分文字。对待切分的字串，从末尾位置开始取最大词长为 m 的字串，如果该字串成功匹配分词字典中的词条，则该字串被认定成功识别一个词，如果匹配分词词典不成功，去掉字串的首字后重新匹配，直至匹配成功或剩余一个单字为止。一次分词操作结束后，将成功切分成词的前一个汉字作为起始，继续重复上述过程。

如果采用逆向最大匹配对“我们是中华人民共和国公民”这个文本进行分词，设最大词长为 7，则分词过程如下：

表 4 逆向最大匹配算法的分词过程

步骤	待切分文本	操作	分词结果
1	我们是中华人民共和国公民	截取七个字符	空
2	人民共和国公民	无匹配结果 字符串去掉一字	空
3	民共和国公民	无匹配结果 字符串去掉一字	空
4	共和国公民	无匹配结果 字符串去掉一字	空
5	和国公民	无匹配结果 字符串去掉一字	空
6	国公民	无匹配结果 字符串去掉一字	空
7	公民	与词典匹配成功， 切分出该词	公民
8	中华人民共和国	与词典匹配成功， 切分出该词	中华人民共和 国\公民
...	...	...	...
	是	与词典匹配成功， 切分出该词	是\中华人民共 和国\公民
	我们	与词典匹配成功， 切分出该词	我们\是\中华 人民共和国\公 民

研究结果表明[25]，逆向最大匹配的切分精准度略高于正向匹配，但这种精准度也不满足实际分词的需要，还需要对分词算法进一步改进。

（三）改进的中文分词算法

正向最大匹配算法在实际应用过程中首先要确定最大词长 length，如果最大词长太短，无法切分出长语句，分词结果被当做新词进行录入，产生歧义，算法的效率较低，切分结果无意义；如果最大词长设置的太长，甚至大于要处理的文本字串的长度，则匹配所花费的时间增加，而且算法的时间复杂度高。因此，本文在传统的正向最大匹配算法的基础上，对分词算法进行改进，改进后的算法无需人为设定最大词长，根据待切分字串的具体情况，在算法运行过程中自动确定[26]。

（1）在处理文本字串的首字时，通过首字 Hash 表对首字的存储位置进行定位，以获

得第一个字的二级索引指针。

（2）将字符串的次字与二级索引的次字 Hash 表进行匹配，若匹配成功，则进入下一步骤；如果匹配失败，则将第一个字作为单字词切分出来。如果待处理的字符串仍大于 1，将切分成功后的下一个汉字作为文本字串的首字，继续进行第一步骤。

（3）找到与次字 Hash 表对应的剩余字符串指针，并获得具有第一个字的前缀和剩余的长度为 n 的字串。

（4）截取第一个字后，根据长字词优先的原则，再匹配剩余字串，如果匹配成功，直接将结果切分出来；如果匹配失败，请删除字符串的最末端一个字并继续匹配其余字串，直到匹配成功为止。如果待处理字符串仍大于 1，则转到第一步骤，否则结束切分。

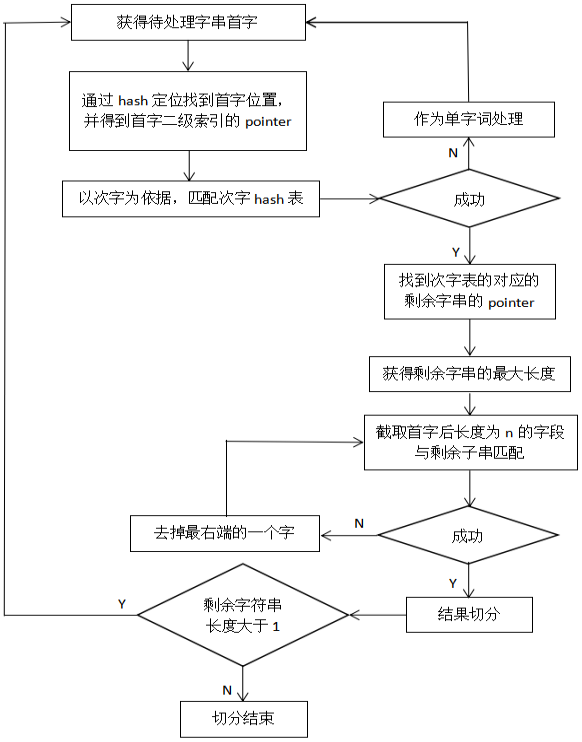

图 16 改进的分词算法流程图

（四）中文分词算法应用到 Lucene 中

1. Token 的结构

通过对 Lucene 的系统结构进行分析了解到 Analysis 是用于各种语言切分词的语言解析器，分词得到关键词和所在域值为组合若干项，在 Analyzer 中项以 Token 的结构出现，它是 Lucene 索引中可直接处理的最小单位，Token 的结构中记录着每个索引字段的信息，其中包括关键词、关键词出现的频次以及关键词出现的位置等信息，在 Analyzer 中项就是分析源文件后返回给检索器的结果。

```
public final class Token
{
String termText; //文本
int startOffset; //开始的位置
int endOffset; //结束的位置
String type = "word"; //文本类型:single 为英文 ASCII 字符或扩展 ASCII
                      //double 为汉字,word 为缺省值
private int positionIncrement = 1;
public Token(String text, int start, int end, String type)
}
```

2.分词模块的文件结构

为了将改进的中文分词算法应用于 Lucene 中，首先需要设计一个继承 Analyzer 的子类，命名为 New；继承[27]Tokenizer 的类 NewTokenizer 在解析处理源文件时重新加载基类方法 TokenStream（），并将分词结果封装到 TokenStream 中。这是中文分词模块的核心部分，能够自动对中英文文本进行区分，对中文文本采用改进的中文分词算法进行分词处理，经过 StopFilter（）处理返回 Token 流。

```
public void New_Analyer extends Analyzer{
public TokenStream tokenStream ( String fieldname,Reader reader){
//使用本文设计的 NewTokenizer 进行分词;
public class NewTokenizer extends Tokenizer{
public NewTokenizer(Reader reader){
dic=Dictionary.load();} //装载词典
public Token next() throws IOException{
//如果结果集为空则返回空值,否则返回第一个词条并从结果集删除;
private ArrayList getTokens() throws IOException{
//将待分析文本读入内存;
ArrayList list=os.analyzerSens(); //使用本文设计的算法进行分词
//将分词结果加入到结果集中;
Return this.tokenList;
```

3.Lucene 中分词处理过程

在 NewTokenizer 函数读入待切分的文本前，先对字符串长度初始化值为 0，确定读入字符的初始位置，在运行过程中先读入一个字符并判断该字符是英文字符还是中文字符，若该字符符合 ASCII 码和扩展 ASCII 码，确定为英文分词，否则采用改进的正向最大匹配算法。在对英文字符进行分词操作时，主要对空格进行定位，如果该字符为空格或其他符

号，则结束位置的值加 1，返回的项以起始位置开始；如果字符不是空格或其他符号，则结束位置的值和字符串长度递增 1 并存储在缓冲区中。在处理汉字时，调用 ChineseParse 函数进行切分，如果字符串数组长度在切分后不为 1,将字符串数组中的首个字符串作为返回的标记。重复该过程得到满足条件的项集。这样无论用户输入的问题语句是中文文本还是中英文混合文本都可以采用正向最大匹配算法进行分词操作。

```
public sealed class NewTokenizer: Tokenizer
{
private int offset = 0;
private int bufferIndex = 0;
private char[] buffer = new char[MAX_WORD_LEN];
public NewTokenizer(TextReader_in)
public override Token Next()
{
int length = 0;
int start = offset;
while(true)
{
char c;
offset+ +;
c = ioBuffer[bufferIndex+ +];
if(( ‘\u0000’ < = c＆＆ c < = ‘\u007F’ )‖ ( ‘\uFF00’ < =c＆＆c< =
‘\uFFEF’ ))
{
 //为 ASCII 或扩展 ASCII,英文处理
}
else
{
 //为汉字，进行基于词典的正向最大匹配分词
}
return new Token(new String(buffer, 0, length), start, start +
length,tokenType);
}
```

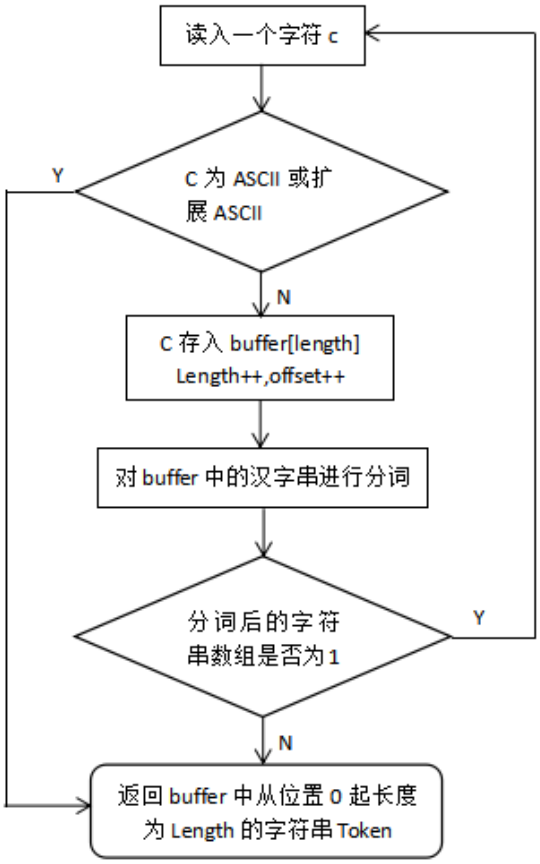

图 17 中文分词处理流程图

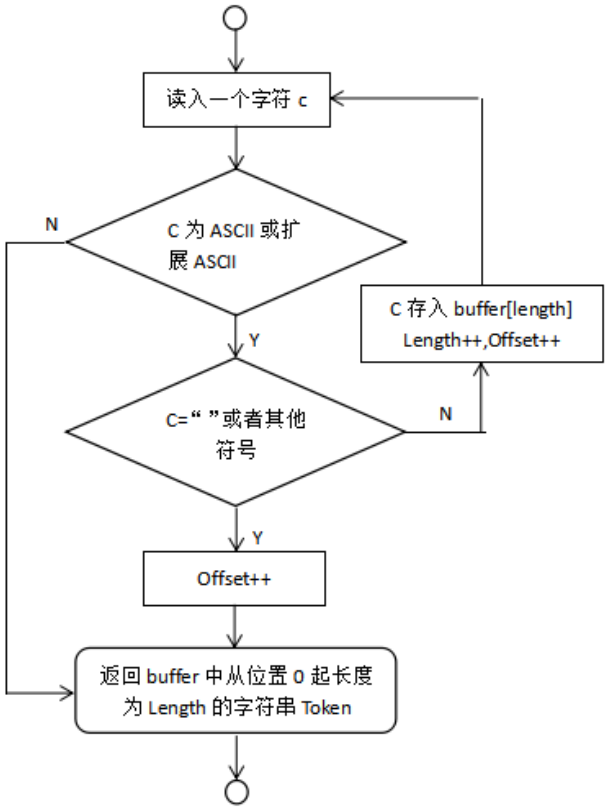

图 18 英文分词处理流程图

四、歧义的发现与消除

（一）歧义的发现

发现歧义的常见方法有双向最大匹配法、逐词扫描最大匹配法、回溯法[28]这三种。

双向最大匹配法通常用于检测算法的切分歧义，比较相同字符串两次的切分结果是否一致，不一致的部分就是歧义字段。双向最大匹配方法的优点在于，仅需要执行一次扫描便找出是否存在交叉类型模糊字段，但识别组合类型模糊字段的能力差。

逐词扫描最大匹配法是使用较多的模糊字段识别方法，并且从字符串的第一个字开始，检测交叉类型模糊字段的准确率更高，选择长度不超过字符串最大长度的子串，匹配分词词典，检查是否出现歧义，若存在歧义按照歧义类型对歧义字段加以标记，若匹配失败，则去掉字串的一个尾字并再次与分词词典相匹配，直至剩余字段的长度为 1，如果该字不是单个字，则从第二个字开始，选择一个长度不超过字符串最大长度的字符串，继续重复上述步骤继续分词。

回溯方法主要用于检测链长为 1 词簇为 2 的交叉类型模糊歧义，与前两种方法进行比较，它减少了与字典比对的次数，优点是算法简单，只要进行一次最大匹配的操作就可以找到歧义。

（二）歧义的消除

歧义消除的能力影响着中文分词的效果。基于规则的分词消歧先标注出切分词的词性，在规则库中进行搜索，根据语法规则查找分词结果；[29]基于统计的分段消歧匹配词典，综合采用最大概率方法和 Viterbi 动态规划的方法，将字符串划分成单个词并计算出所有分词结果的概率，以其中最高概率作为切分结果；从训练语料库中学习基于实例的分词消歧，并结合语境中的词性、句法关系和词语搭配等各种知识进行消歧，其核心是使用特征值计算实例之间的相似性，并选择与消歧结果最相似的训练实例。消除模糊歧义的常用方法包括互信息概率统计方法、t 检验方法和双字耦合度方法等。[30]郑晓刚等人在一种组合型中文分词方法中对交集型歧义字段中链长的分布进行了研究，研究结果表明，链长为 1 的歧义字段出现的次数占全部歧义的 53%左右,其中出现交集型歧义的字段数占全部的 60%以上，对于链长为 2 的歧义字段出现的次数占全部歧义的 42%，其中出现交集型歧义的字段数占全部字段数的 36%以上，上述的这些歧义字段占总歧义字段的 95%以上，若能实现对其中大部分歧义的消除，对中文分词的促进有着不可估量的意义。

从中文的构词特点来看，词是由相邻的字构成的组合，字与字在文本中共同出现的频率越高，构成一个词的可能性越大，通过比较互信息值的大小，对相邻字符的紧密程度进行判断。因此本文采用双向最大匹配法发现交叉模糊字段，并利用互信息来消除歧义[31]。互信息的消歧策略是在交叉点模糊字段处比较前一个字符串与后一个字符串的互信息的大小。对于形如 XJY 这样的歧义字段，若 M（X，J）>M（J，Y），将 XJ 作为一个词切分出来，否则将 JY 作为一个词切分出来。M(X，Y)是互信息的值，用来体现字符 X、Y 的紧密程度，则有公式：

$$M(X，Y)=\log_2(P(X，Y)/P(X)*P(Y))$$

P(X，Y)=n（X，Y）/N　　P（X）=n（X）/N　　P（Y）=n（Y）/N

用 n（X，Y）表示字符 XY 在文档中出现的次数，n（X）、n（Y）分别表示字符 X、Y 在文档中出现的次数，N 表示整个文档中字符的个数，P(X，Y)表示 X、Y 两字相邻共现的概率，P(X)、P(Y)分别表示字符 X、Y 在文档中出现的概率。

第六章 自动答疑系统的实现与性能测试

一、开发环境与工具

本文自动答疑系统采用 B/S 体系结构，使用 Tomcat 9.0 作为 WEB 应用服务器，部署 Myeclipse 进行开发，网页设计采用 Dreamweaver CS4，生成动态交互的 WEB 服务器应用程序，后台数据库为 SQL Server 2008，脚本语言采用 JavaScript，操作系统为 Windows 7。

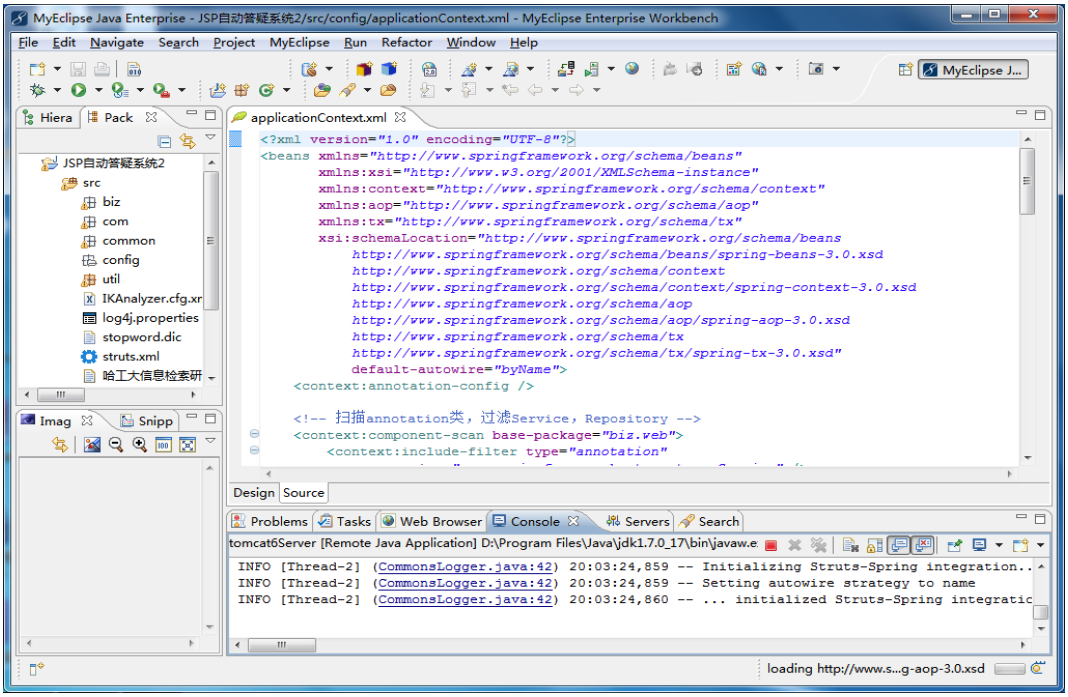

图 19 Myeclipse 界面

二、 主要功能模块的实现

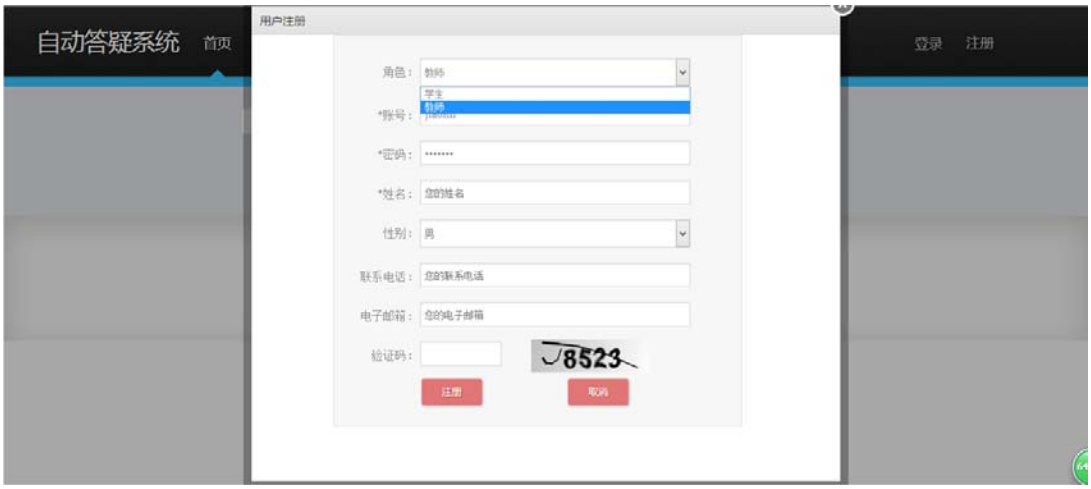

图 20 用户注册页面

1. 系统登录模块。登录前注册用户信息，输入账号、密码、姓名、性别、联系方式和电子邮箱地址，并选择相应的角色，不同的用户拥有不同的权限和功能模块，只有注册成功的合法用户才能登录系统。

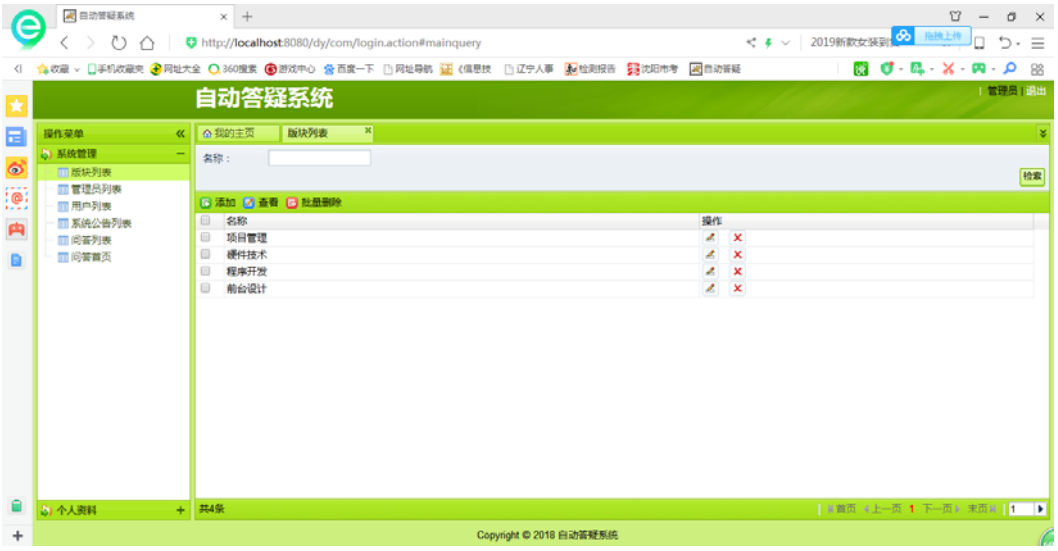

图 21 管理员界面

2. 知识库的组织。知识库信息数据的来源之一是在系统建立之初，由领域专家和教师根据教学大纲和实际教学经验从书本、网络等多种途径整理出该领域的常见问题和答案，构建成为系统初步的知识库；另一来源是在系统运行过程中，教师对学生疑难问题的解答对知识库的扩充，以及日常维护操作，如根据学生提出的问题，教师对系统添加删除和更改知识库等。

图 22 知识库的组织

3. 自动答疑模块。自动答疑是学生用户的核心功能模块。学生用户可以在显示的输入框中输入关键词或者自然语言进行提问，系统通过对问题文本切分来提取关键词，并使用 Lucene 全文搜索引擎在知识库中进行搜索。如果知识库中存在相应的问题，系统会向用户返回正确而简洁的答案，并突出显示与学生提问相同的关键字，方便学生点击查看；若知识库中无类似问题，系统将未解决的问题统一转交由教师进行回答。学生用户之间也可以

对问题进行探讨，学生通过快速提问可以将疑难问题发布在问题列表中，等待其他同学的回答。

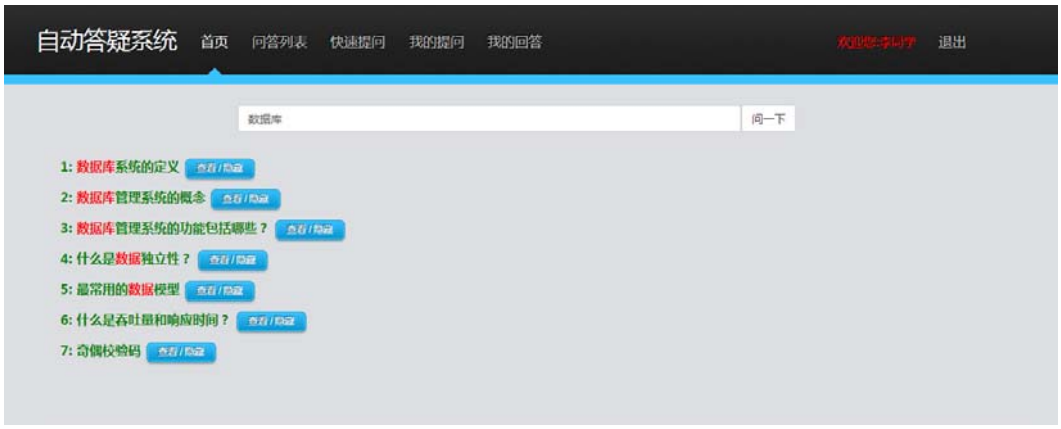

图 23 关键词提问

图 24 自然语言提问

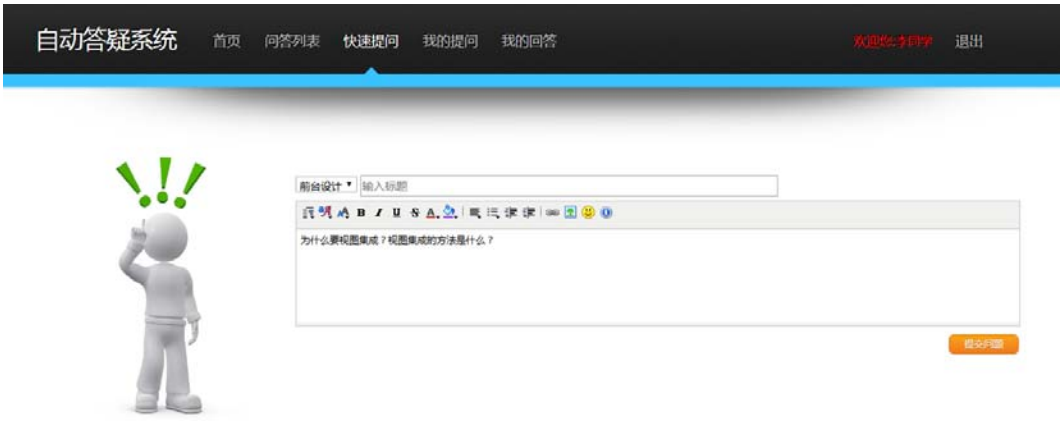

图 25 学生用户快速提问界面

4. 教师答疑。教师的主要功能模块是查看要解决的问题列表，并回答系统中未解决的问题，教师回答的问题和答案会自动添加到知识库中以更新知识库。教师点击要回答的问

题，系统将教师已解答的问题通过系统公告的形式呈现给学生。教师也可以单击添加按钮，完成新问题答案的录入，对知识库进行补充和更新。

图 26 教师答疑界面

三、 性能测试分析

在本文中，自动答疑系统的性能测试主要针对答疑模块，该模块根据检索答案的准确率和检索速度方面进行测试。以《数据库系统原理》课程中的相关知识作为知识库的测试样本，样本数量为 1300 条左右。经过系统的实际运行得到的测试结果如下表所示：

表 5 系统准确率测试结果

学生 ID	提问个数	正确个数	正确率	提问方式
20181001	16	13	81%	单个关键词
20181002	21	19	90%	多个关键词
20181003	22	21	95%	多个关键词
20181004	19	16	84%	自然语言
20181005	20	19	95%	自然语言
20181006	32	28	87%	自然语言
20181007	18	16	88%	自然语言
20181008	22	25	88%	自然语言
20181009	21	19	85%	自然语言
20181010	23	20	86%	自然语言
平均正确率			88%	

表 6 系统检索速度测试结果

编号	问题	所用时间（毫秒）
1	常用的基于磁盘的技术策略是什么	70
2	SQL 中谓词 EXISIS 用于测试什么集合	3
3	在关系模型中,键是由什么组成	5
4	在关系模型中,记录称为	4
5	数据独立性最高的是什么	7
6	模式/外模式映像用于解决数据的	6
7	关系模型的三元完整性约束是什么	6
8	数据操纵语言 DML 分为哪几类	4
9	关系代数、基本运算、5 种	3
10	后援副本、用途	3
平均时间	11.1	

从测试的结果来看，自动答疑系统的平均正确率达到 88%，具有良好的运行性能，能够满足用户正常的答疑要求。当用户使用单个关键词进行搜索时，正确率不高，这是因为问题太广泛。多个问题中都包含相同的关键词，从而导致呈现给用户较多的检索结果，学生需要在搜索结果中执行第二次检索，以找到自己真正需要的答案。在用户具有良好的提取关键词的能力时，用户也可以使用多个关键词进行查询，这时检索的速度可以得到明显的提高，但是值得注意的是，并不是所有的问题都能够采用关键词查询的方式进行检索，当遇到复杂的问题时，学生很难从冗长的题目中提取出关键词，更多时候学生所使用的关键词并不能很好的表达问题的实际意义，在这种情况下，采用自然语言提问的优势就显露出来了。自动答疑系统的平均检索速度达到 11.1 毫秒，从表中发现系统第一次检索的时间最长，这是由于文件的 I/O 操作以及 Lucene 在检索缓存方面优化的原因，这样做可以保证之后的问题检索响应速度很快。从性能测试结果来看，该自动答疑系统具有良好的运行性能， 可以满足学生的日常提问需求。

第七章 结论

一、研究总结

通过对国内外自动应答系统研究现状的比较，分析了目前中文答疑系统的现状和存在的问题。根据需求分析，设计了系统结构，自动答疑系统选用 Lucene 全文检索技术作为系统的检索工具，确保了答疑系统的查全率和检索效率。针对 Lucene 全文搜索引擎在具有强大检索能力的同时也存在着致命的弱点，Lucene 中没有成熟的中分分词模块且自带分词器的分词精度不达标这一问题，基于 Lucene 的系统体系结构，本文提出了一种基于双字哈希索引的词典结构，减少了待切分文本与词典的匹配次数，以此提高匹配效率。根据使用范围、操作难易等多方面的考量，选择正向最大匹配算法作为系统的中文分词算法，并对该算法进行了改进，改进后的分词法不再需要人为设定最大词长，能够在运行的过程中自动地获取最大词长，以此改善了答疑系统的精准度，它对交叉型模糊歧义也具有一定的消歧能力。经过系统的性能测试，该自动答疑系统的平均正确率达到 88%，具有良好的实用性，可以满足学生日常的答疑需求。随着计算机相关技术领域研究的愈加深入，自动答疑系统在查全查准率方面将会更加完善。在不久的将来，自动答疑系统能够为学生的学习提供更加便捷的服务和有效的指导，为教师的教学提供更有价值的参考。

二、未来展望

本文在知识库的建立等方面还有可提升的空间，在系统建立之初，需要由领域专家和教师根据多种途径整理出该领域的常见问题和答案，构建成为系统初步的知识库，当知识库出现空缺时，也需要领域专家教师对知识库进行手动补充，知识库不可能自动从网络资源中检索答案并补充知识库。此外，还可以建立文本分类模型对知识库中的问题进行自动分类，使杂乱的问题形成体系，减小候选答案，索引和二级检索策略可以对答案进行快速定位，提高系统效率。本文对交集型歧义有一定的消歧能力，但缺乏对未登录词的识别。在未来的工作中,将收集更多的测试数据来验证系统的分词准确率，进一步优化系统的相关性能，这些都是下一步要研究方向。自动答疑系统中的中文分词技术不仅广泛应用于搜索引擎中，而且还广泛应用于自然语言理解和信息处理领域。它也不断拓展于机器翻译、自动分类、简繁体转换、分类校对、自动摘要等多方面。它的应用也不仅仅局限于文本输入的模式，生活中无论是微信聊天中的语音与文字间的转换，还是语音导航时输入的语言，甚至是智能家居的普及都与中文语义的理解有着莫大的关系，众多的电子产品要想体现出智能性，也离不开中文分词技术。此外，对语言文字的处理，包括句型识别、词频统计、词结构分析等问题也有帮助，中文语义的理解与我们的生活密不可分，渗透在日常生活的方方面面，只有不断的提高中文分词技术的性能才能更好的应用于社会服务于大众。

参考文献

[1] 耿立伟.答疑系统在网络远程教育中的应用[J].信息与电脑，2017（14）:72-73

[2]马新意,王剑辉.自动答疑系统中文分词模块的设计与实现[J].信息技术与信息化，2019（01）：19-22

[3]孟旭生.改进的中文分词算法在自动答疑系统中的应用研究[D].大连交通大学，2008

[4]王明东.基于 WEB 的自动答疑系统的研究与实现[D].华侨大学，2012

[5]季永华，许华虎，沈敏.自动答疑系统的研究与实现[J].计算机工程与应用，2005（05）：27-29

[6]王丛林.在线自动答疑系统设计与开发的研究[D].东北师范大学，2010

[7]尹炳龙.消除交叉歧义中文分词算法的研究与应用[EB/OL].（2017-01-08）[2019-02-28].http://ishare.iask.sina.com.cn/f/av5k1Aa2ob6.html

[8]吴代文，杨方琦.Lucene 在数据库全文检索中的性能研究[J].微计算机应用，2011（06）:53-59

[9]zhanghefu.全文检索系统与 Lucene 简介[EB/OL].（2017-03-27）[2019-02-28].https://blog.csdn.net/zhanghefu/article/details/1542230

[10]李伟.基于用户兴趣模型的新闻自动推荐系统[D].复旦大学，2009

[11]songylwq.web 信息模糊检索等 Lucene 的实现[EB/OL].（2011-02-21）[2019-03-07].https://blog.csdn.net/songylwq/article/details/6197140

[12]王璐璐，袁毓林.走向深度学习和多种技术融合的中文信息处理[J]. 苏州大学学报， 2016（04）：160-167

[13]林鹏祥.微博信息检索系统研究与开发[D].华中师范大学，2014

[14]夏天，黄文，马骏涛，李光伟.Lucene 全文检索软件及其在学科信息服务平台中的应用[J].图书情报工作，2011（21）：106-109

[15]张献力.互联网网页蕴含高动态交通信息的实时搜索与语义理解技术研究[D].浙江工业大学，2014

[16]王学松.Lucene+nutch 搜索引擎开发[M].北京：人民邮电出版社,2008

[17]吴虎子.中文网页获取及自动分类技术研究[D].武汉理工大学，2007

[18]何恒飞.主观题智能阅卷的关键技术研究[D].北京工业大学，2013

[19]杨敏利.高职单片机知识问答学习支持系统研究[D].山西师范大学，2017

[20]牛力，王久为，黄蕊.面向政府决策的知识库建设问题研究[J].档案学通讯，2015（04）:56-60

[21]彭文颖.知识库在呼叫中心的应用[J]. 无线互联科技，2012（07）：176

[22]向晖.DRIS 系统中的中文自动分词模块设计与实现[D].华中科技大学,2007

[23]向晖，郭一平，王亮.基于 Lucene 的中文字典分词模块的设计与实现[J].现代图书情报技术,2006（08）:46-50

[24]孔维亭,闫宏印.基于 Lucene 的自动答疑系统的设计[J].电脑开发与应用，2012（04）：32-34

[25]付敏.一个改进的中文分词算法及其在Lucene中的应用[D].华中科技大学，2010

[26]王瑞雷，栾静，潘晓花.一种改进的中文分词正向最大匹配算法[J].计算机应用与软件.2011，28（03）：42-44

[27]李庆虎，陈玉健，孙家广.一种中文分词词典新机制——双字哈希机制[J].中文信息学报.2003（04）：13-18

[28]宁可为．自动答疑系统知识库文本的索引研究[J]．电脑知识与技术，2009（35）:10100-10101

[29]康晨阳.基于避免交集型歧义的最大匹配算法改进的研究与设计[D]．西南电子科技大学，2012

[30]江耿豪．自动答疑系统中文自动分词模块设计与实现[J]．现代计算机，2010（02）：8-10

[31]郑晓刚，韩立新，白书奎.一种组合型中文分词方法[J].计算机应用与软件，2012（07）：10-16

个人简介

本人概况

姓名：马新意
性别：女
出生年月：1993 年 11 月 17 日
籍贯：辽宁省沈阳市

教育背景
2012 年 9 月-2016 年 6 月 就读于沈阳师范大学 计算机科学与技术（师范）专业 学士
2016 年 9 月-2019 年 6 月 就读于沈阳师范大学 计算机应用技术专业 硕士

攻读硕士期间发表论文
《自动答疑系统中文分词模块的设计与实现》发表于信息技术与信息化 2019 年第一期

致谢

时光荏苒，岁月如梭，三年的研究生时光转瞬即逝，回顾这三年的学习生活，我有过迷茫，有过辛苦付出，但更多的是收获和感恩。

首先，我要感谢我的导师王剑辉教授，王老师学识渊博、治学严谨对我产生了深远的影响。在学术上，王老师对我悉心教导使我学到了很多专业知识和技能，加强了我对计算机应用相关知识的理解和掌握，他不仅培养我分析问题、解决问题的能力，还在实践中教会了我很多为人处事的道理，对我产生了潜移默化的影响。王老师对我的影响不仅在学术上，更蕴含在平时研究生生活的点点滴滴中，云山苍苍，江水泱泱，先生之风，山高水长，能够成为王老师的学生，让我深感荣幸。其次，我要感谢沈阳师范大学数学与系统科学学院的其他老师，感谢邓立国、杨姝、张岩、周传生老师对我学习和生活的帮助，感谢您们在科研和未来人生道路上对我的鼓励和祝福，虽然平时见面的机会不算很多，但是每一次相见时您们的关心和问候我都感激在心，衷心地感谢您们对我的关心与栽培。此外，我还要感谢学哥、学姐在论文写作期间对我的辅导和关心，感谢我的室友和一起同窗，共同奋斗学习过的同学对我的帮助，三年的研究生生活因为有了你们而变得更加精彩充实、闪闪发光。

最后，我最要感谢一直都关心我、支持我、教导我、不辞辛苦地养育了我二十几年的父母，谢谢您们在生活中对我无微不至的照顾、给予了我无尽的动力，正是由于您们的支持、付出与理解才让我得以安心学习，让我以积极阳光的态度顺利完成了三年的学术生涯，没有任何后顾之忧。在我落笔的这一刻，我不禁思绪万千，纸短情长，它记载了我三年的青春时光，承载了这三年来的我坚持不懈的努力，字里行间中皆蕴含着深厚的情谊。