中图分类号：TP315　　密级：公开
UDC：004　　学校代码：10082

河北科技大学

HEBEI UNIVERSITY OF SCIENCE AND TECHNOLOGY

硕士学位论文

教学网站智能答疑系统设计与实现

论 文 作 者：李印鹏
指 导 教 师：张冬雯　教授
企业指导教师：敦少博　高级工程师
申请学位类别：工程硕士（在职培养）
学 科 、领 域：计算机技术
所 在 单 位：信息科学与工程学院
答 辩 日 期：2018 年 5 月

河北科技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。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。除文中已经注明引用的内容外，本论文不包含任何其他个人或集体已经发表或撰写过的作品或成果。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：郭帅　　指导教师签名：张之雯

2018年5月29日　　2018年5月29日

河北科技大学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，同意学校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。本人授权河北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。

本学位论文属于

√□不保密。

（请在以上方框内打"√"）

学位论文作者签名：郭帅　　指导教师签名：张之雯

2018年5月29日　　2018年5月29日

Classified Index：TP315 **Secrecy Rate**：Publicized

UDC：004 **University Code**：10082

Hebei University of Science and Technology

Dissertation for the Master Degree

Design and Implementation of Intelligent Question Answering System for Teaching Website

Candidate：Li Yinpeng

Supervisor: Prof. Zhang Dongwen

Enterprise Supervisor ：Senior Engineer Dun Shaobo

Academic Degree Applied for：Master of Engineering

Speciality：Computer Technology

Employer：School of Information Science and Engineering

Date of Oral Examination: May, 2018

摘 要

目前，网络教学的教学模式逐渐兴起。但是，在众多的网络教学课程中，主要存在两方面问题，其一，老师只是单一讲解，很难实现学生的个体化需求，对学生提出的问题不能一一解答；其二，教师可能在某一重复的问题上浪费了极多的时间来进行解答。所以，如果能有一款系统可以实现对单一学生提出问题自动解答，上面的问题就迎刃而解。

本文结合教学网站的具体需求，提出了智能答疑系统的解决方案和开发了一套基于 B/S 架构的智能答疑系统。首先，本文详细研究了答疑系统的相关技术以及相应的理论，从而确定对答疑系统的功能需求；其次，分析并研究了目前的中文分词算法，并最后确定使用逆向最大匹配算法来实现系统的功能；然后，研究了语句相似度计算方法，语句相似度由词形相似度、语句长度相似度、词序相似度三个方面决定的。本系统根据这三个指标来对学生输入的问题进行语句相似度的匹配；最后，利用上述研究的内容，开发了一套基于 Web 界面，采用 Java EE 技术，Apache tomcat 作为网站服务器，SQL Server 2012 作为后台数据库的智能答疑系统。在尽可能低成本、高效率情况下，去满足教学网站智能答疑的业务需求。

基于课程教学网站的智能答疑系统的研究目标在于方便学生学习，使学习者能够在教师不参与的前提下，利用智能答疑系统迅速准确的得到知识点的解答，从而最大程度的利用网络提升自学效率。同时，智能答疑系统直接替代了传统的教师人工对某一单一问题反复解答，极大的减轻了教师劳动强度。除此之外，智能答疑系统利用了大数据时代特点，对学生提出的问题加以归纳，重点、难点以数据化形式加以体现，在一定程度上提高了教师的课程教学质量。

关键词 教学网站；智能答疑；分词；语句相似度

Abstract

At present, the network teaching mode is emerging gradually. However, in many network teaching courses, there are mainly two problems, one is that the teacher is only a single interpretation, and it is difficult to meet the students' individual requirements, the questions from students cannot be answered one by one; The other one is that teachers may waste a lot of time on a repetitive question. Therefore, if there is a system that can answer questions from a single student automatically, the problem will be solved.

For the specific requirements of the teaching website, this thesis puts forward the solution of the intelligent question answering system and develops a set of intelligent question answering system based on the B/S architecture. Firstly, the relevant theories and technologies were studied, and the demands of the related functions of the system were ascertained. Secondly, the existing Chinese word segmentation algorithm is analyzed and discussed, the reverse maximum matching algorithm based on string matching is chosen to realize the functions; Then, the semantic similarity calculation method is studied, and the semantic similarity is determined by the similarity degree of the word, the similarity of sentence length and the similarity of word order. According to these three indicators, the system matches the semantic similarity of students' input. Finally, the intelligent question answering system which takes Apache tomcat as a Server and SQL Server 2012 as the database is developed by Java EE technology. The system meets the demand of the intelligent answering of the teaching website with low cost and high efficiency.

The goal of the research on the course teaching website based intelligent question answering system is to facilitate students' learning, enable learners to use the intelligent question answering system to get knowledge quickly and accurately without the teachers' participation. Thus the efficiency of self-study is improved with the best use of the network. At the same time, the intelligent answering system directly replaces the traditional mode that teachers answer single problems repeatedly, and greatly alleviates the labor intensity of teachers. In addition, intelligent question answering system, which uses the characteristics of big data era, can summarize the questions from students and reflect the key points and difficulties in a digital form. This can improve the teachers' teaching quality to some extent.

Key words Teaching website; Intelligent answering; Participle; Sentence similarity

目 录

第1章 绪 论

1.1 论文研究背景和意义

1.1.1 研究背景

如今，随着信息技术的不断变革更新，获取知识的方式变得不再单一化。移动互联网的诞生，使得生活方式开始打破时间、地域的限制，教学过程也有了新的产物－网络教学[1]，从此，知识的获取不再是单一的在课堂上获得。在全球范围看，网络教学也在成为一项热门的教学方式，因其方便灵活而备受追捧，并衍生一大批职业。在线智能答疑系统将协助网络教学模式，实现教育的新型改革，提高广大国民的素质与自身能力。当然，网络教学处于初步发展的阶段，作为一种时代的新兴产物，还是有许多不足与改进的空间。当前的网络教学存在这种情况：学生如果想在网上向老师进行请教，往往会面临着老师没有及时在线而无法及时答复的情况，久而久之，这种问题就会被忽略，从而学生心中的疑惑也就越来越多。因此，需要一种能够及时对学生问题进行反馈的网络系统。

在互联网广泛应用的今天，检索可以在一定程度上实现人们对于知识定向化的需求。但是分析目前主流搜索引擎存在很多相同的缺点。首先，传统搜索引擎相关性很大，输入问题后，检索出的网页几十个上百个，用户不能快速的获取想要的信息；其次，人的主观因素，在检索时人们常用几个简单的关键词来进行问题的检索，但经常出现词不达意的问题，这时候引擎自然也不会对于用户的问题给出准确的解答。最后，以关键词为基础的索引、匹配算法在一定程度上有所进步，但是这个算法更加注重了每组独立的关键词，无法实现相关几个关键词的排列组合的语义，检索效果差强人意。

上个世纪六十年代，人工智能逐渐崭露头角，有很多研究人员就提出能否让计算机通过自然语言来快速回答人们的问题，这就是指：智能答疑系统[2]。二十一世纪以来，网络和信息技术呈现蓬勃发展势头，人们想更快地获取信息的愿望重新促进了智能答疑系统的发展，智能答疑系统自然而然成为了当前新型化信息化教育发展的焦点。在课程教学网站[3]中应该如何更好地使用也是一个正在探讨的问题。因此，如何设计与开发一个实用的智能答疑系统是当前网络教学的需要解决的重要问题。

1.1.2 选题意义

本文以课程教学网站中搭载的智能答疑体统的设计与实现为主要研究对象。该系统将学习者的自主性与讲师的指导性有机结合，使得学习过程人性化，避免了传统的死板教学模式。智能答疑系统的研究意义是实现时时点对点的学习，跳出传统

的空间、时间、地域对与学习的限制。学生提出自己的问题之后，系统利用自然语言处理技术[4]对检索内容进行语句分析，与原有的知识题库[5]进行快速匹配，最后实现对问题的准确回答。

利用智能答疑系统[6]迅速准确的得到知识点的解答，从而最大程度的利用网络提升自学效率。同时，智能答疑系统直接替代了传统的教师人工对某一单一问题反复解答，极大的减轻了在线教师劳动强度。除此之外，智能答疑系统利用了大数据时代特点，对学生提出的问题加以归纳，重点、难点以数据化形式加以体现，在一定程度上提高了教师的网络教学质量。

1.2 国内外研究现状

经历了一波又一波互联网技术的革新，从 WEB1.0 到 WEB2.0 等的改变，信息化浪潮进入了新篇章。这种新的变化正在将当下的生活方式推向新的变革，例如数字经济为人们带来电视观看的新享受、数字经济为经济社会的发展带来新机遇等，这一切的变革将在现在即未来给人们带来生活上的便利和更佳的选择。社会大趋势对互联网教育的号召，推动互联网教育得到进一步的发展。

网络技术的日新月异给教育带来了明显的改变，之前的改变更多集中在使教育不在受地域和时间的限制，如身在国内也可以接收到国外大家的课程教学，同时可以反复听讲，任何时间想听就可以听，时间不在是教育的限制因素。当然，这只是之前的变革，现今还要求在开展教育的过程中增加教的内容、学习的吸引以及对教育与学习的评价等环节，使教学成为一个新的链条。因此，在这种新的模式下，学习和教育不在是像以前一样单纯的存在于教师开展教学、学生在网络的一端接受学习、最后简单的给予一个反馈的教学过程，而是开始成为一个新的模式，发展成为一个新的教与学的过程。现代互联网教育[7]即运用当下发达的计算机和互联网技术，将教师在课程当中的授课音频、授课视频通过实时或者存在时间差的方式将其传达到网络环境中，进而校园内外的学生均可以接受到这样的教育。与以往通过纸质资料进行知识传输的函授教育以及通过广播电视进行知识授课的广播电视教育不同，当下互联网教育的传达方式是结合现在先进的多媒体、电脑及网络技术等将知识整合传送，离开任何一部分都难以实现传送，这种教育之所以能够有效的开展得益于现今计算机和互联网技术的迅猛发展，同样它是教育界的一次创新和革新，离不开通信技术与信息处理技术的发展。专家指出，互联网为教育带来的变革是教育界的一次创新，是对教育资源的一次有效整合，可以更好地实现教育资源的优化配置，给更多人一个教授教育的机会，是未来教育的新趋势和新发展。

互联网教育[8]目前有优势所在，同时也存在一定的问题。其主要的优势是其资源丰富，它可以容纳来自全国甚至全世界各地高校企业等教育内容，同时可以给学生提供一个互动的教学环境，即学生在听课的过程中可以及时跟老师进行反馈和提问，

同时授课老师也可以及时给予回复，在这样的教学环境中学生的积极性和主动性受网络教育这种对方的刺激，进而得到极大的提高。但是，互联网教育是将教育网络化，但目前信息化、网络化的发展还处在发展的初期，未来还有极大的更新空间，这也就极大的限制了其发展，主要体现在硬件和网络服务上，同时由于传输速度不是特别快，使得实时传输教育始终受限，进而影响教育的效果。因此，需要一种智能的答疑系统来弥补这种不足。

人工智能在 60 年代的时候处于刚刚起步阶段时，就有研究者曾经提出是否可以让计算机回答人们的一些问题，这种回答必须要用自然语言（所谓自然语言即人们都能懂得语言，通常是为理解而产生的各种语言，例如中文、英语、法语等语种），这被称为智能答疑系统。智能答疑系统 IQAS（Intelligent Question Answering System）[9]是一种计算机的程序，通过这个程序用户可以肆意运用各种语种提出问题，计算机会根据大家的提问进行回答，并且尽可能用见解和准确的语言进行回答，并有效的实现人机交互。随着社会和网路技术的推进，加之社会的变革速度让人们逐渐意识到知识的重要性，并且希望得到及时的答复，为此这就推动了该系统的发展。进而使该系统成为信息化教育发展的热点。

为更好地让读者了解该系统，我们将从下面从国内国外的研究现状来进行详细介绍：

随着慕课在中国的发展，各有关部门都在进行统计，根据教育部统计结果显示，慕课在我国正处于高速发展阶段，在国内我国一些水平较高的学校已经建立自己的慕课平台，这样的平台大概有十余个，同时一些国内的高校也开始在国际一些慕课平台上开设课程，进而将中国教育推向世界。数据显示，截止目前大概有 3200 多门课入住一些线上慕课平台，这些课程多数来自于高校，源自我国大约 460 多所高校。在这些平台上不仅有高校学生在学习还有一些社会热爱学习的人，数据显示这些人数已经高达 5500 万人次。这样庞大的开课和上课群体使得我国的我国慕课数量成功位居世界第一的位置。教育作为一个细化群体，除了目前的大家熟悉的一些免费通识类教育（如网易公开课等平台）之外，也逐渐开始对各个阶段的教育进行细分。有些平台开始专注于基础教育，即幼儿园到高中；也有平台专注于线上培训，如英语类培训、公务员考试培训、一些证书类培训等。中国慕课的迅猛发展。使得国外极其关注，来自国外数据研究调研显示，截止到 2017 年，中国在线教育市场规模预计达到 1941 亿元。基于这种发展规模，未来中国的在线教育将呈现更加迅猛的发展形势。

这种发展虽然迅猛，但是却有分析者发现，在我国的这些平台中真正运用到智能答疑这种技术还是很少。因为现在我国的一些平台在答疑设置上更多的是在平台中运存一个留言板以供学习者提问，可见这种答疑方式还是集中在人际交互方式，

在应用于技术的方面还是需要进行下一步的分析和研究。目前国内的此类系统，一般有两种方式，一种是人工答疑，即人工在线回答；一种是智能答疑，即通过信息化技术进行分析问题和回答问题。但是对于人工答疑来讲，目前存在的最大的问题是在给予用户回答的时候缺乏对问题进一步深入的分析和研究，相对来讲最大的问题就是内容控制难，整个设计结构不合理。

随着 Internet 技术的不断发展和国内互联网资源的不断完善，中国互联网上涌现了许多网络教育系统。比如：以北京大学的网络教育系统下的智能答疑系统为例，该校在系统中给同学们提供讨论和答疑等方式，主要通过论坛的形式开展。为了给学生提供更加便利的答疑环境，该校在系统中专门设置了一些讨论区，这些讨论区是根据不同主题开展了，例如基础课程、数学、语言类、管理类等等。这样学生可以更快的寻找自己要答疑的区域，提出问题，等待系统的回答。同时，还可以看到该区域中其他的问题，如果有些问题中已经有自己疑问的问题了，就可以直接采纳，避免等待。还有一个是北京师范大学的在线远程教育平台，叫 Vdass[10]。这套系统的一个优势就是它有一个自有的知识库，该知识库中存储了过往学生提出的质疑和对应的解答，教师在设计教案、设计课程的时候可以有效的参考其中的内容，以便更好地为学生提供教学。同时，也为学生提供了一个平台，当学生存在疑问的时候，可以通过搜索的方式在历史库中寻找问题及解答，这样可以更快的解决疑问，实现答疑的自动化。

相比国内的这类系统，国外的此类系统则显示出更多的独立性，因为它们更多的是独立运行，简单讲就是不同于国内一些高校或者机构创建自己的平台，国外更多的是不属于任何教学或者机构的，是一个独立的系统[11]。另一个不同于国内的是，国外的系统更多像是在提供问答的资源，有问题的学习者更多的是在平台上提出某些问题，而不是像国内一样在平台上寻找一个完整的教学程序，国外的提供者更多的是针对一些问题一部分进行提问。有的是基于学习者对学科的兴趣，有的基于提问者在学习和工作中存在的问题。因此，国外的此类系统在设计的过程中更加偏向于简洁性，即他们在设计系统的时候更多的是针对答疑这项功能展开。也因此，它们的答疑系统中主要的是电子邮件和信息板，关于聊天或者讨论的模块比较少，有些甚至没有。即使这样，国外的智能答疑系统有一些具有其特殊的特色。例如，国外有个系统叫 Ask Jeeves for Kids[12]，这是一个重点在于为提问者提供问答的系统，这个提问中，不仅支持自然语言，而且还提供了一个交互的平台，通过这个交互系统可以系统可以深入理解和分析用户所提出的问题，并通过深入分析问题，给予提供准确的分析，确保问题回答的精准。

除此以外，国外还有一些具有代表性的系统，这些系统在某些知识领域可以给提问者带来答疑，此类系统有：

1) 美国 Ask Jeeves 公司开发的 Ask Jeeves 系统[13]，该系统的特性在于它的交流性。当提问者在平台上提供问题时，该系统有一个类似于人工智能交流的功能，它可以跟提问者进行交流，通过交流进一步确认提问者对问题的理解和疑问之处，进而明确获悉提问者的疑问。然后根据本系统中的智能库进行问题寻找，进而给提问者提供明确的答案。唯一的一个缺陷是，该系统中目前给提问者回答的结果还是以网页的形式出现，而并非提问者所需的直接答案。

2) 还有一个叫 Answer Bus[14]。相对其他系统来讲，这个系统比较成熟。该系统主要基于 Alta Vista、Google、yahoo News、和 Wisenut 等多种重要搜索引擎技术支持的智能答疑系统。该系统中，当有提问者提出并且提交了问题之后，系统会根据其问题中的语句进行检索，其检索过程中不在按照单一的语句，而是按照多个语句检索，以便更好地为提问者回答。

1.3 研究主要内容

本论文研究的是课程教学网站智能答疑系统的设计与实现。其核心工作是研究设计智能答疑系统的整体结构构成，创建对应的模块知识库，设计提问者提出的问题的分析算法以及快速匹配算法，基于以上内容实现系统的功能计算。

1) 充分调查教学网站智能答疑系统的现状和存在的问题，实际调研分析系统工作过程中的实际工作流程，从而明确智能答疑系统的功能和需求。

2) 找寻关键点。对于这类系统来讲，关键点的一步是怎样准确理解用户所提出的问题，这决定了后期对答案的匹配。而我们使用的手段就是对问句进行分词[15]。目前有很多分词的算法，如，逐字匹配算法，最小匹配算法(Minimum Matching)等。本文中所提出系统拟采用逆向最大匹配法[16]进行切分词语处理。而对于在当前系统无法进行分词解答的问题，系统将自动转为人工答疑，即系统将问题转到等待教师回答的状态。

3) 准确无误的寻找答案，是另一个该系统设计的重点所在。通常来讲，在系统中，一句话中会有多个关键词，每个词对应不同的点，如果分析和对应有误，那么最后给出的答案极有可能严重背离提问者的问题。本文主要通过分析词型、语句长度和词序三个方面，来进行语句相似度[17]的计算。

4) 针对教学网站智能答疑系统的相关需求，利用上述研究的内容，基于 B/S 架构模式[18]下，采用 Java EE 技术，使用 My Eclipse 作为开发工具，Apache tomcat 作为网站服务器，SQL Server 2012 作为后台数据库，构建起智能答疑系统[19]。根据用户提出的问题能够快速准确地返回答案，满足智能答疑的需求。

1.4 论文组织结构

第 1 章　介绍了本文的研究背景和选题意义，并对实现教学网站智能答疑系统

的研究现状进行了调研，最后对本文的研究内容进行了总结。

第 2 章　对相关技术进行了说明，包括了中文分词算法，语句相似度的计算，以及本系统采用的 Java EE 的平台技术、服务器，数据库等。

第 3 章　是对教学网站智能答疑系统的需求分析。主要包括系统功能上和非功能需求。具体包括系统的总体框架和各功能模块上的需求，以及系统中的角色划分，以及对数据库的需求分析。

第 4 章　开展调研。重点调研该系统中的组织结构和运作流程，以便后期可以更好地开展设计，本文重点在对模块功能和数据库的设计。

第 5 章　详细介绍了系统关键技术的实现，主要包括中文分词处理技术和语句相似度匹配算法技术。

第 6 章　详细介绍了开发工具及环境，以及展示系统界面设计及功能。最后，描述了系统测试环节，说明该系统可以正常运行，满足用户基本需求。

最后是对全文内容总结，并提出改进意见。

第 2 章　相关技术概述

2.1　中文分词算法

中文分词算法[20]主要是切分汉字，切分的方法是根据序列进行切分成单独的词，通常是一个一个的，把这单独的词通过空格与英文字符进行分开，中文字符在语义识别时，需要把数个字符组合成词，才能表达出真正的含义。分词算法是文本挖掘的基础，通常应用于自然语言处理、搜索引擎、智能推荐等领域。

目前关于分词算法主要有四种类别，第一类是基于字符串匹配的分词方法，需要一定策略开展；第二类是基于理解的分词方法，需要运用模拟的方式；第三类是基于统计的分词方法，需要考量其出现频率和计算方法；第四类是基于规则的分词方法，需要分析汉字串和计算机存储中的关系。

基于字符串匹配的分词方法[21]更多的是运用一种匹配方法，即将需要分析的字符串或者汉字内容与现有机器中的词库进行对照和匹配，若在库中找到对应的内容，则匹配成功，这样就找到了对应词条。基于此类方法的特征，有几种不同的分析、匹配方法。第一种，如果扫描方向存在不同，匹配时会用到一种正向和逆向的匹配方式，即从两个端点开始匹配数据。第二种，如果所给数据长短不同，此时会启用最长或者最短匹配方法，以便更加快捷的实现对接。第三种是取决于所给定的内容中有没有特殊的标记，如果存在标记就使用分词与标注相结合的方法，如果不存在标注就使用单纯分词方法。目前最常用的几种方法是，第一正向最大匹配法，根据含义不难看出，是跟多数人的方向感一样，从左向右的方式开展数据的匹配。第二种是逆向最大匹配法，即是跟正向相反的方向，是由右向左的方向进行寻找。第三种是最少切分法，即从给定的数据中切除部分数据，寻找词数最小的内容。

基于理解的分词方法[22]，更多的是考量计算机对语句的理解程度，通过机器的识别让词语更好地被理解和识别。这种方法的主体思路是理解，通过对给定的内容进行分析、分词，然后结合数据库进行词句、语义、语言的分析和处理，然后在排除歧义的内容。一般它由分词、句法语义两个子系统和总控系统构成。对歧义词句的判断和排除就是在总系统的指导下开展的。一般这种分词的方式是基于大量语言和信息模块知识点的基础上，才能更好地实现。众所周知，汉语的语言系统是极其复杂的，每个词在不同的环境会存在多种释义，因此会给极其阅读和分析带来极大的困难，为此这种方法目前一直处于试验的过程中。

汉语中，形式上两个字在一个文章中出现的频率越高、且位置越接近，说明它们越有可能构成一个词，而且在文章中是稳定的词，代表的意思在此文中也是相对固定的。此时就用到一个词叫互现，即看词语出现的频率和程度。基于统计的分词

方法[23]就是运用到了这个互现信息，根据互现信息中例如汉字 A 和 B 一共出现了多少次，最终计算的就是这它们的频率。由于这种方式的特性不用跟之前的方式一样进行切分词语，因此，有人跟它叫做无词典分词法或统计取词方法。任何一种方法都是有其优势，也有其存在的问题，这种方法最大的缺陷在于，有时提取出的高频出现的词句，不是生活中或者专业上的组成，简单讲是一种机器的组成词，这会影响数据的准确性。

基于规则的分词方法[24]主要是通过分析和匹配的方式实现，即根据系统中设定的程序和方式将需要分析的内容与计算机中的内容进行匹配，如果能够得到对应的数据则匹配成功，下一步在分析其意义。运用此方法的过程中会用到的方法有三种，第一种是最小匹配算法(Minimum Matching)，即从给定的要分析的数据或者字符串的左边开始进行比较，通常会先比较最开始出现的两个词，将其跟机器数据库中的内容匹配，然后如果数据库中正好存在，则对该词进行分词，然后进行下一个数据的匹配，若没找到对应值，则终止。第二种是最终正向（逆向）最大匹配法(Maximum Matching)，从名称中可以看出这其中包含两种方式，一种是正向的最大匹配方法，一种是逆向的最大匹配方法。这两种方法的区别在于方向上是一种相反的。为了更好地理解这种方式，我们就以正向的方法为例进行简要分析。首先要有一个假设的前提，即计算机机器中当下最大可以包含的汉字有 N 个，然后假设从需要分析和匹配的数据中取前 N 个作为分析的主体。假使在对应的过程中，从计算机库中找到该词，则首度寻找是成功的，这时可以将这个单独的词分开。之后需要继续后的信息匹配，在匹配则从带分析内容的 N+1 处开始取前 N 个作为分析的主体，具体方法同之前一样，最终全部切分则匹配结束。第三种是逐字匹配算法，此方法的核心机制是树型词典机制。更直观的理解就是把它看成一个树，先从树的某一个分支开始，一步步寻找。这种方法最大的优点就是速度快，但是存在的阻碍是构造一棵树和对树进行维护是非常复杂的。为此为了更好地进行匹配和分析，一种新的算法就是集中几种方式于一体。

2.2 语句相似度计算

在用词上，在哪使用，在什么时候使用，和哪些词一起使用。也就是说，如果人们要进行有意义的交流，比如人们在分析、探究某个物体的时候，一般会涉及到一个情景，在这个特定的情境中该事物具备特有的意义，此时就可以根据其特有情境进行语义分析。目前在技术和分析领域来看，对不同的词语进行语义和语境的分析具有特有的意义，有时会借助一些技术方法进行实现，例如本体[25]知识库构建、机器翻译、构造特有知识库等。

现今在运用的过程中，根据语句相似程度开展运算功能的方法，主要有两种。第一种是大家熟悉的语义词典[26]，这种方法在实施过程中主要是将概念组成一棵树

形结构，然后根据结构进行推算；还有一种是根据这个语句中词语进行分析，通过统计多种答案，最终进行分析，得出唯一解。

之前提到了根据词句上下文的概念进行推算是基于语料库[27]的词语相似度研究中常用的而一种方法。之所以开展此方法，是基于大家一个共同的认知，即一段文字中的上下文的释义可以给其提供一个判断的依据，而且是比较准确的判断。目前在使用这类方法中，最常用的方式就是词语向量空间模型[28]。

有一本书《同义词词林》[29]，这是梅家驹等人于 1983 年编著的，虽然时间有些久远，但这部词典中最大的功绩在于他不仅介绍某一个词语的意义，还将与该次有关的其他内容进行介绍。唯一的遗憾是，该书后期没有根据时间的推移逐渐进行词语内容的更新，后来为了研究和时代的需要，有高校学者根据该书进行了扩展，编著了《同义词词林扩展版》。在这本新的词典中，总共有大约 7 万条词语，并且最大的一个创新点是它根据词义进行了分析和编排。由于时代的需要，后期为了更好地实现数据匹配，许多计算机等信息库中都进行了同义词的匹配。本文基于同义词识别的需要，引入了《同义词词林》，作为教学网站的智能答疑系统的语义知识资源库[30]。

2.3　开发工具调研

2.3.1　开发语言及平台

Java 2 是一个广泛应用的语言版本，该语言是由 Sun 公司为了解决在开发复杂的应用时通常会出现应用之间的多样性，同时不同应用之间没法交互的问题，为了解决这一问题，Sun 公司提出了一种不同企业之间对应用程序进行开发时需要遵循的一个开发标准。在该标准成功解决这些问题，并不断被更多的人所采用后，Sun 公司又进一步对 Java 语言的标准进行了改进和完善，在这次改进过程中，主要提出了容器、组件、以及分层等概念来扩展 Java 语言的功能，从而形成了 J2EE（Java 2 Platform, Enterprise Edition）平台技术[31]。

J2EE 是一种体系结构，是一个平台，是一个服务系统。作为体系结构，目前他可以为企业提供计算和运行环境用于开发和部署多层体系的调用。作为一个平台，目前多数企业在开展数据分析的过程中都会用到他，基本上具备了工业的标准。作为一个服务系统，主要体现在它可以给企业提供多种计算服务。在使用 J2EE 进行系统开发时，可以实现在很大程度上降低对不同应用上开发的成本，从而使得开发的速度更快，效率更高，另一方面，通过使用 J2EE 作为一个统一的开发平台，还能够保证开发出来的系统的稳定性和安全性。J2EE 的开发框架是一个开放的框架，因此不需要进行付费，这也就为资金不是很充裕的小型企业来说，提供了便利，该类企业可以通过使用该框架进行统一的开发与管理，同时还可以获得较好的开发质量和

性能，而不需要投入过多的开发成本。

J2EE 目前有几大优势。第一，目前它具备标准版的主要功能，例如数据存取非常方便、支持 Internet 应用中使用并可以保护数据等。第二，它还支持 Java Servlets API、EJB（Enterprise JavaBeans）、JSP(Java Server Pages)以及 XML 技术。J2EE 体系结中有一个中间层集成框架，通过这个框架可以满足一些高要求的应用需求，比如费用低、可用性高并支持扩展的需求。这种平台开发中实现的统一性提供，极大的降低了 J2EE 在开发多层应用的费用和复杂性，并且为现有的一些新的程序的开发带来了有力支持。

虽然 J2EE 很先进和实用，但是随着技术的进步，总是会有新技术出现。例如 Sun 公司后来基于 J2EE 又研发并上市了 Java EE（Java Platform，Enterprise Edition）[32]，并且用它取代 J2EE。这项新的技术而后成为企业开发和应用的新的标准。但由于之前编写的 JSP 代码中，因为存在混淆，主要是显示代码和业务的主要逻辑上，这种混淆导致数据之间的不嵌套，导致程序在后期的维护和扩展上存在巨大的问题。为了程序的易维护性和可扩展性，这就需要我们使用 Java EE 技术来进行项目开发。

My Eclipse 企业级工作平台（My Eclipse Enterprise Workbench ，简称 My Eclipse），它是 Java 语言开发工具，它可以实现支持 JSP，CSS，Java script，Spring，SQL 等。

My Eclipse 作为一种插件的集合，对开发 Java, J2EE 来讲是十分必要的。同时，这款插件还具备较强的功能，同时适用范围比较广。本文基于 Java EE 平台技术，使用 My Eclipse 作为开发工具开发了教学网站的智能答疑系统。

2.3.2 服务器技术

JSP 的发布，后期需要新的引擎进行跟进，为此之后出现了新的 JSP 引擎。Apache Group 一直致力于服务器的研究和开发，它们在成功研发并上市了 GNUJSP1.0 之后，继续进行研究，而后出现了 Tomcat[33]。它是 Sun 公司官方推荐的 Servlet 和 JSP 容器。另外 Tomcat 由于它技术先进、性能稳定，为此可以为任何人提供从因特网上免费下载的服务。当然业界对两者的评价虽然都很高，但是如果这两者能够有效地结合，才是最佳架构。

本文中智能答疑系统是基于 Web 界面，采用 B/S 架构模式, 所使用的网站服务器是由 Apache tomcat 提供的。此 Web 服务器是的代码是免费面向公众的，属于轻量级别的，主要应用在中小型系统和并发访问用户不是很多的场合下，是开发和调试 JSP 程序的首选。实际上 Tomcat 是 Apache 服务器的扩展，但运行时它是独立运行的，所以当你运行 Tomcat 时，它实际上作为一个与 Apache 独立的进程单独运行的。基于本系统所采用的 Java EE 的平台技术，Tomcat 能够更好的去提供一个 Web 应用的界面，同样也提供了一个成熟的管理工具。

2.3.3　数据库技术

新的数据库技术随着社会的进步仍然在不断出现，Microsoft SQL Server 2012 是微软公司根据技术发展的需要研发的新产品，为更好的将服务提供到位，它设计两个版本，一个是为企业服务，一个是面对大众的标准版。SQL Server 2012 在数据管理服务中，提供了全面的数据库管理系统。在关系数据库中，其功能是非常强大的，一方面可以进行结构化数据类型，一方面还保障了在存贮中安全性，与此同时还能轻松创建高性能的数据库应用程序，是新时期技术发展的要求。在为企业级用户提供服务的过程中，将几项功能进行整合，可以更好地给企业提供服务。

SQL Server 2012 在一般情况下能够实现对 OLTP[34]和 OLAP[35]这两种数据库的管理。

(1)　OLTP 数据库　一般数据库中也会存在一些多余信息，为了更好地实现数据库的更新和对接，该数据库中特意将数据被组织存放到关系表中。这样就方便了 SQL Server 2012 的开展，使得该系统可以容纳大量用户开展事务的处理，同时还可以更改 OLTP 数据库中的实时数据，促进数据的更新。

(2)　OLAP 数据库　这个数据库主要是利用 OLAP 技术来实现对海量数据的组织和汇总，这样有助于后续的分析程序实现对数据的快速响应和分析，有时可以实现和确定实时的数据结果。SQL Server 2012 中提供了专门的 Analysis Services，用于对数据库中各类不同数据的组织、管理和分析，从而适用于大量数据下的分析服务，从而便于通过数据分析进行决策支持等功能的实现。

SQL Server 2012 在两种安全级别上验证用户：登录身份验证和对数据库用户账户和角色的许可权限。

登录身份验证：如果用户准备建立与 SQL Server 2012 的连接，就必须拥有相应的登录账户。SQL Server 2012 包括 Windows 和 SQL Server 两类的登录身份验证机制，两种身份验证方式都有不同类型的登录账户。

许可权限验证：在每个数据库内部，您可以为用户账户和角色分配执行（或限制）某些动作的许可权限。在用户成功访问了数据库之后，SQL Server 2012 就可以接受命令。

2.4　本章小结

本章主要对教学网站智能答疑系统使用的相关技术进行了详细介绍，包括中文分词算法，语句相似度算法和开发此系统所用到的相关工具，为后续系统的实现奠定了基础。

第 3 章　智能答疑系统需求分析

需求分析[36]是一个分析软件成型前得关键步骤。虽然开发人员对于计算机软件的编程相当熟悉，但是并不了解用户的业务领域与相关需求，这个差异如果处理不好可能直接在功能性上造成软件开发的失败。在软件开发的前期，开发人必须与用户进行相关概念与相关功能要求的确定，最终形成一个完整的、清晰的、一致的需求说明。需求分析对于一款功能性软件的开发周期至关重要。

3.1　系统功能需求

3.1.1　系统总体框架

智能答疑系统[37]需要实现三方面的功能。首先，系统需要借助计算机网络系统，实现最基本网络在线答疑功能；其次，在该系统的界面需要有最直接的检索、生成以及后台软件的相关管理功能；最后，需要系统提供延时性的解答功能，在检索者没有在题库中检索到自己需要的答案时，教师在登录时，可以进行补充解答，并实现数据库题库的更新。数据库的题库是智能答疑系统中的核心组成部分，它存储所有的问题和答案。具体流程见下图：

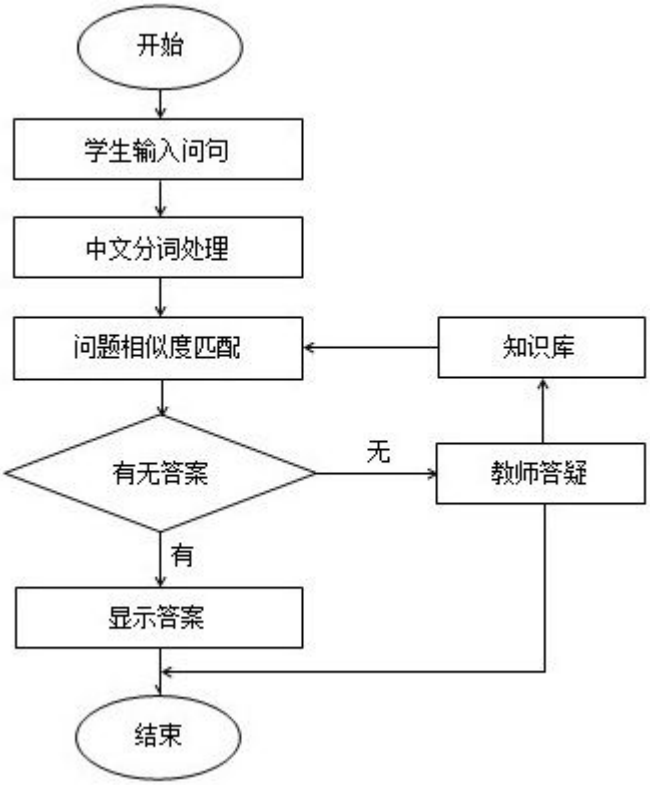

图 3-1　系统总体流程图

系统工作流程[38]是对输入的中文问句进行分词，之后需要与知识库中的知识进

行匹配，并根据算法计算得到相似度最高的问题，并将其对应的答案作为智能答疑的结果返回给用户。具体实现中，对问题语句进行中文分词处理[39]，搜索知识库，将问题语句的分词段与知识库检索当前问题语句分词段求解语句相似度[40]，并记录结果；当完成对知识库的检索时，求出语句相似度最高的一条问题，并获取该问题对应的知识答案，作为本次智能答疑问题的答案。系统主要的功能是智能答疑，但同时应当实现智能性化的服务，学生在使用时对于常见的问题，在检索中输入相关的关键词，即可完成答案的检索生成。考虑到，有些问题，关键词信息量较少，检索可能没有结果或者多个结果，所以系统又加入了自然语言处理技术，即学生可以使用自然的语言对问题进行详细描述。系统依然可以识别问题，然后能够生成让用户可以接受的、简短的、有效的并且正确的答案。

3.1.2 角色分析

根据本系统的总体框架，对本系统的用户的角色进行分析，总结出如下几类用户的角色：分别包括对老师、学生和系统管理员三类，其中学生用户可以输入要查询的问题，如果系统可以智能回答则显示相应的答案，否则如果学生输入当前系统中无法回答的的问题，系统自动则提交到数据库，等待老师的回答。教师用户主要对学生用户提出的系统无法自动回答的问题进行答疑。系统管理员主要对系统进行维护，并对系统的权限进行管理。其中各角色之间的关系如图 3-2 所示：

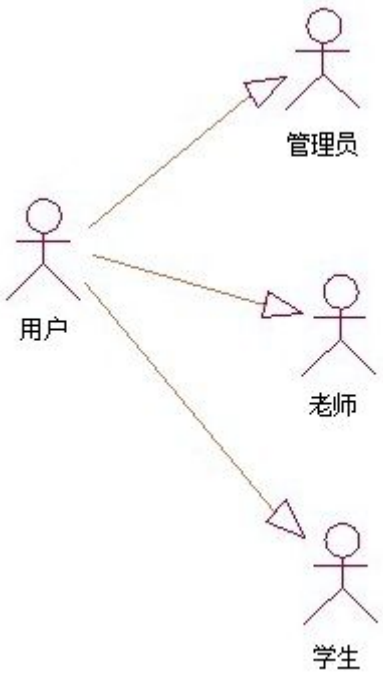

图 3-2　系统中的用户角色

教师角色的主要功能包括注册登录，查看当前待回答的问题，回答学生提交的

问题，并提交数据库几部分的功能。教师用户的用例图如图 3-3 所示。

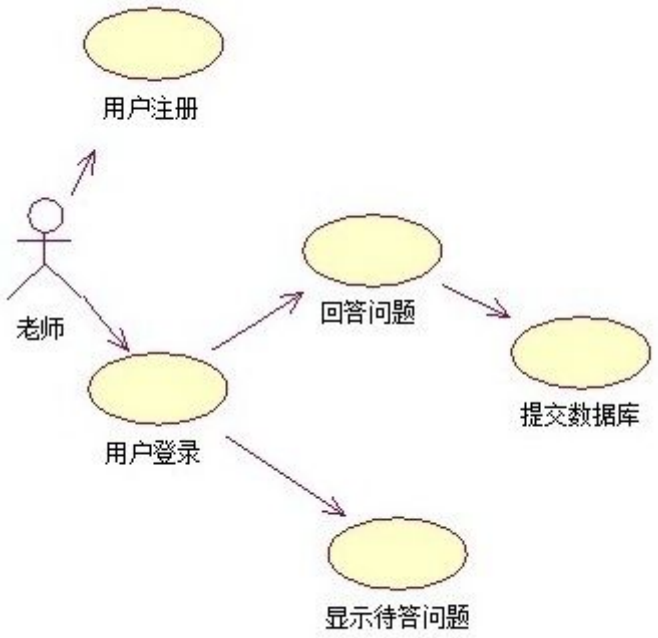

图 3-3　教师用例图

学生角色的主要功能包括账号注册，登陆，查看问题答案，查询问题答案，以及把问题提交到数据库等几部分的功能。学生用户的用例图如图 3-4 所示：

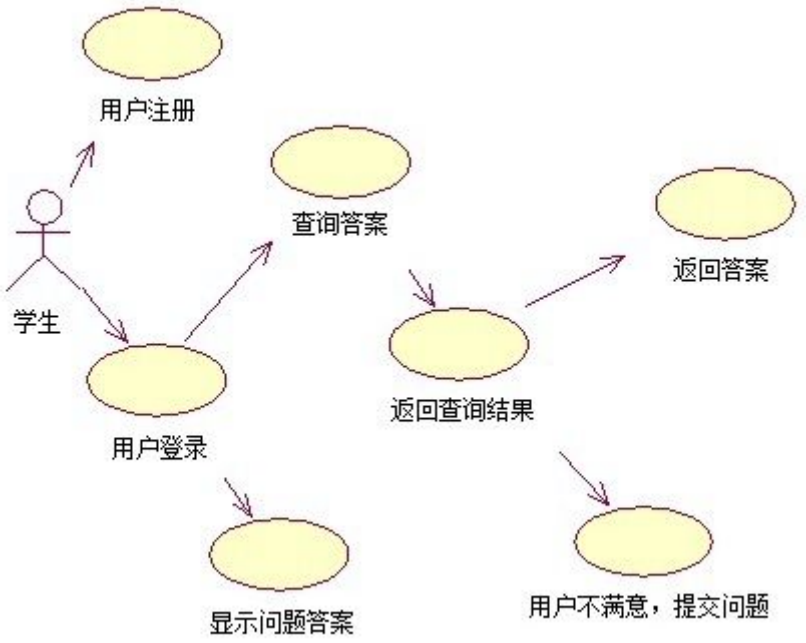

图 3-4　学生用例图

3.1.3 各部分功能需求

根据教学网站智能答疑系统中的工作流程[41]，我们将该系统根据三类用户对该系统中网站的功能进行划分：这三部分的功能主要包括教师模块的功能，学生模块的功能，以及系统管理员模块对应的功能。其中，每个用户类型对应的功能的用例图和详细介绍如下所示：

(1) 教师用户对应功能需求 教师用户类型对应的功能需求主要包括用户注册，用户登录，查看待回答区需要回答的问题列表，以及对这些待回答的问题进行回答，然后把教师回答过的问题提交到知识库中，便于下次学生提问时对类似的问题进行网站的智能自动回答。

教师用户实现功能如图 3-5 所示：

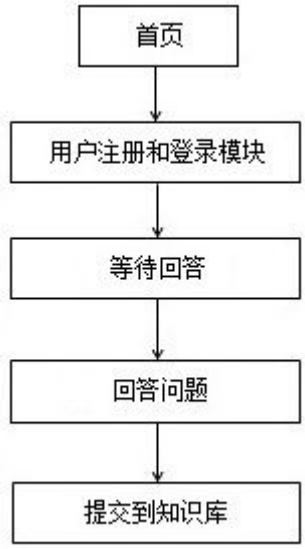

图 3-5 教师用户类型需要实现的功能

其中用户的注册和登录的功能主要是对自己的基本信息进行填写，提交后经过系统管理员的审核，从而完成教师用户的注册，教师用户只有登录后才有权对系统中的问题进行回答，该模块同时还实现了用户的登录和错误验证。教师用户在实现登录后，教师可以在自己的工作区查看不同学生提交的当前系统无法回答的问题，老师可以对此模块对待答问题进行回答，回答后的问题连同答案一起被提交到数据库，从而便于学生下次再次对相同或者相似的问题进行提问时，系统自动生成已经存在数据库中的相关问题的答案。

(2) 学生用户对应功能需求 学生用户类型对应的功能需求主要包括用户注册，用户登录，查看典型问题答案，查询问题答案，以及把系统当前未能给出满意答案的问题提交到老师待回答的数据库中。

学生用户实现功能如图 3-6 所示：

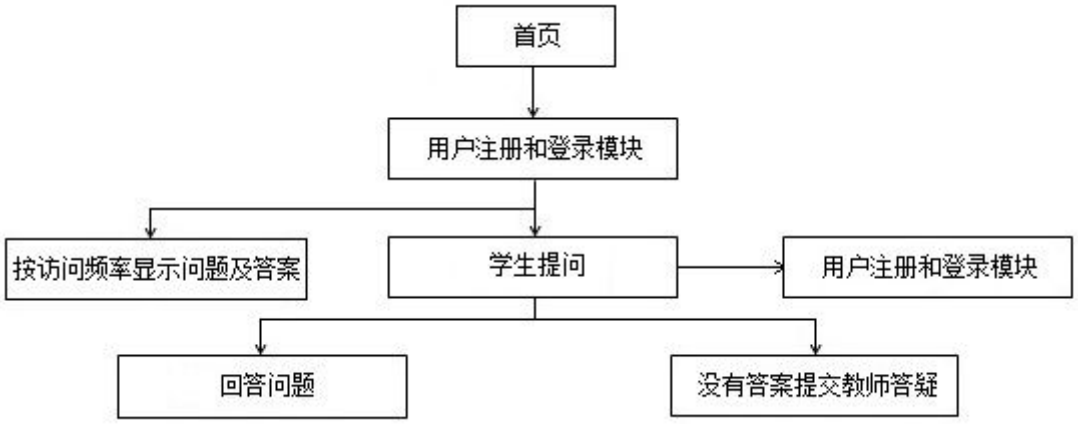

图 3-6　学生用户类型需要实现的功能

其中用户的注册和登录的功能主要是对自己的基本信息进行填写，提交后经过系统管理员的审核，从而完成学生用户的注册，学生用户只有登录后才有权查询问题和提问题。学生用户登录到课程网站后，课程网站将会在学生主页上显示学生容易犯错的典型问题和答案，并按照学生访问的频率进行排序，将学生查看越多的问题放在常见问题列表中，供学生查看。除了这些已有的问题的答案外，学生还可以根据需要自己输入需要查询的问题，系统根据词法分析，识别学生用户提问的问题特征，然后再数据库中查找相关问题的答案，将相关性最高的问题答案显示给用户，如果当前数据库中没有学生需要查询的问题的答案，那么将学生提交的问题放到教师的待回答区，等待教师对该问题进行人工回答。

(3)　管理员用户对应功能需求　管理员用户类型对应的功能需求主要是对系统中的用户进行管理，主要包括账号资料的维护，用户管理。系统管理员可以对用户的账户进行删除，对有关账户的属性可以进行更新。

管理员用户实现功能如图 3-7 所示：

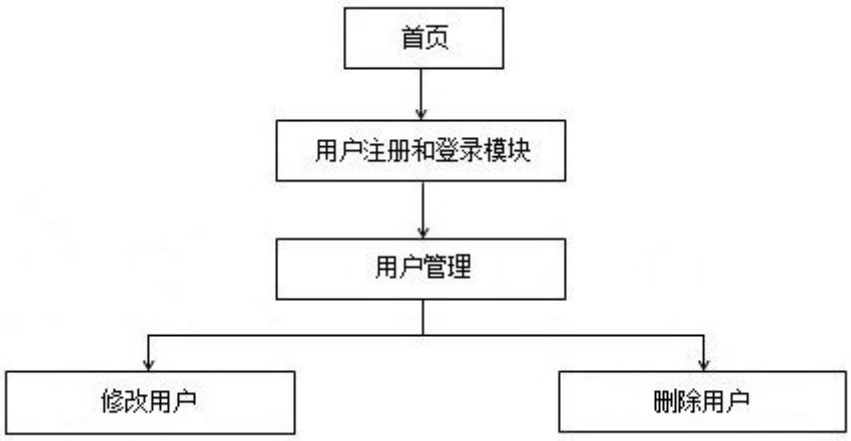

图 3-7　管理员用户类型需要实现的功能

3.2 系统性能需求

对于这个智能答疑系统[42]要求，我们的设计理念为不仅要实现智能答疑的基本作用，而且尽可能的体现系统的人文需要，使得学生在学习的时候不会感到枯燥，从而实现系统更大范围的推广应用。

因此除了实现以上的主要功能外，我们同时考虑网络系统性能、应用系统性能和数据性能等不同非功能性属性的需求。

1) 网络性能方面，要求数据传输网络畅通、快捷、安全，应具有高可靠性、可扩展性、可管理的能力；

2) 答疑系统应当能满足答疑自动解题的基本功能，在实现过程中平稳，具有较高的可靠度。

3) 系统的整体界面应当美观，检索、输入等位置清晰可见，图形与表格比例合理使得用户观看舒适，系统以后使用中应当维护较为方便，具有一定的可开发性。

4) 系统的设计技术采用结构化，系统的设计成果应当具有良好的可移植性，在新的需求时，可以对系统进行扩充。

3.3 本章小结

本章主要从智能答疑系统所实现的功能进行了需求分析，分别从系统总体架构，系统所涉及到的角色，各部分实现的功能和系统性能，这几个方面进行了需求的说明，为下一章系统设计奠定了基础。

第 4 章　智能答疑系统总体设计

4.1　概要设计

教学网站智能答疑系统[43]由学生用户模块、教师用户模块和管理员用户模块三部分组成。对学生用户来说，首先进行用户注册，其次登录系统，进入系统后会显示问答列表，我的问题和我要提问三个选项模块。当学生点击我的问题模块，显示该学生所提问的所有问题和答案；学生可以通过我要提问模块提出问题等待教师或者系统给出答案。对教师用户而言，同样先注册在登录系统，教师用户包括问答列表和等待回答两个选项模块。教师通过等待回答模块来对学生提出的系统无法回答的问题，进行回答；问答列表显示所有访问过系统的用户所提出的问题及答案。管理员用户模块包括用户管理和删除用户两个选项模块。系统总体功能模块如图 4-1 所示：

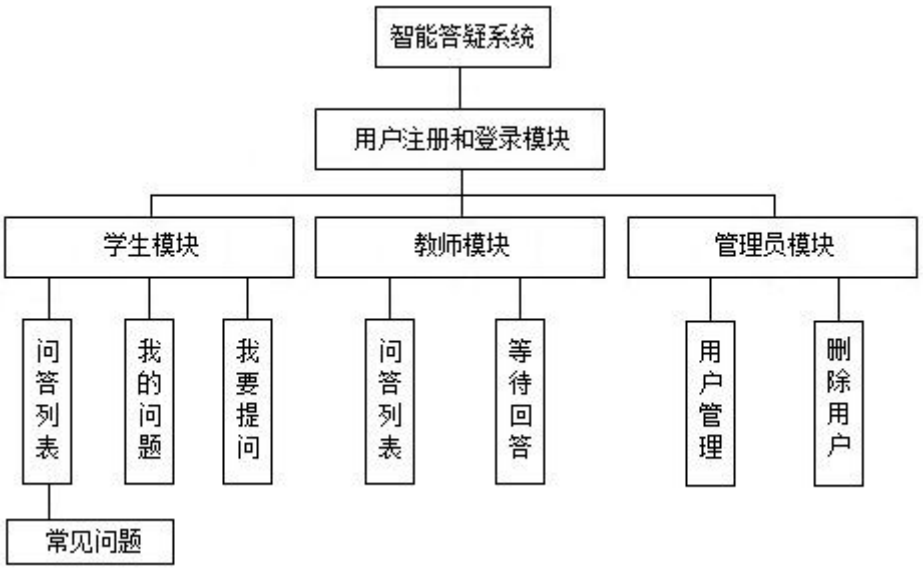

图 4-1　系统总体功能模块图

4.2　模块功能设计

本系统采用模块设计方法[44]进行设计。模块法使得系统的结构层次分明，对于前期开发过程的管理与后期代码的调试工作具有良好的效果，这也为以后系统的稳定性与扩展性提供了可能。与此同时，在系统使用过程中，如出现问题，可以根据模块划分快速找到问题所在。系统的模块化工作需注意以下特点。

1) 整个系统划分的模块数据尽量小，划分出来的单个模块彼此应当尽量互不干涉，在以后的维修中，修改某一模块代码产生的连锁修改降到最低。

2) 各模块彼此之间的联系不宜复杂化，连接接口编程尽量简单。

4.2.1　用户注册和登录模块

(1)　用户注册　用户在使用该系统进行答疑之前，首先需要在该答疑网站上进

行用户注册。用户的注册过程是首先登录该网站，然后在网站上点击用户注册进入注册页面。在这里需要输入用户的相关信息，在输入信息的过程中还会对这些信息进行检查，如果这些信息有误的话，那么将反馈给用户进行重新的信息输入，若注册信息皆填写无误则将该用户的信息插入到数据库中，并且显示注册成功。用户注册过程的活动图如图 4-2 所示：

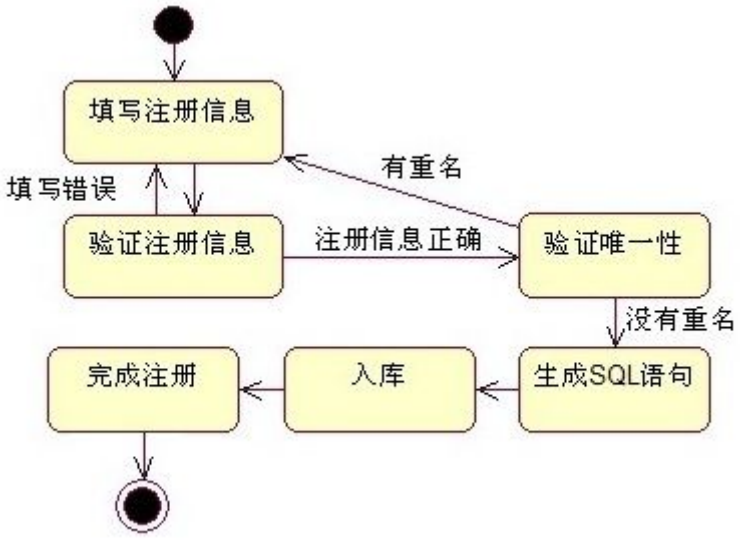

图 4-2　用户注册过程活动图

(2)　用户登录　当用户注册成功后，就可以采用注册的信息进行用户登录。用户登录过程的活动图如图 4-3。首先填写用户的登录信息（包含用户名和密码），系统以用户名为关键字读取数据库用户表中数据，系统会根据登录信息与数据库中信息进行一一对应实现信息匹配。如果实现匹配的完整性，则系统数据库中的用户权限也将实现，之后系统反馈到前台的即为用户成功登陆，下一步的页面也会随之跳出。否则的话，则在用户登录页面显示“输入用户名或密码无效，请重新输入。”

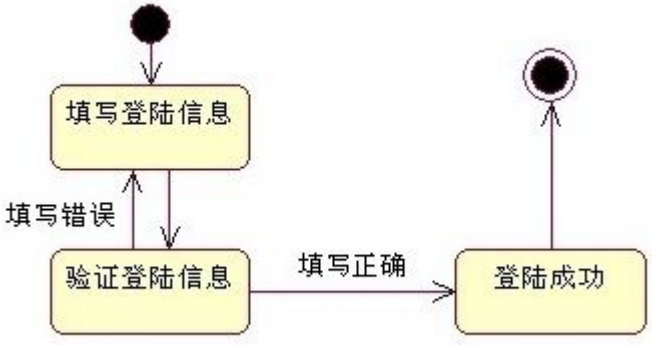

图 4-3　用户登录过程活动图

4.2.2　学生用户模块

学生用户模块主要包括问答列表、常见问题、我要提问和我的问题四个子功能

模块。

(1)　我的问题　学生登录系统后，可以通过点击我的问题按钮查看该学生提问过的所有问题，信息均显示在页面上。流程图如图 4-4 所示：

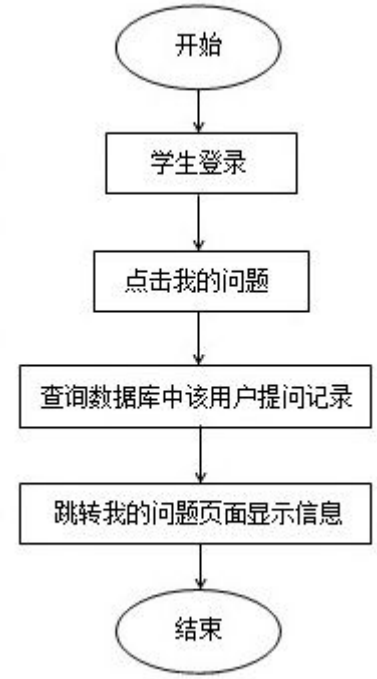

图 4-4　我的问题流程图

(2)　我要提问　该模块是智能答疑系统关键技术所在，学生可通过我要提问提出问题并获取问题答案。

我要提问模块的主要流程：

1) 学生登录系统后，在学生首页左上角点击我要提问按钮；

2) 点击后，在页面上输入问题内容，并点击提交；

3) 学生点击提交按钮后，系统将问题请求转到后台，并在知识库中匹配相似问题；

4) 若匹配到最为相似的问题，则匹配成功，系统返回问题的答案显示在页面上；

5) 若系统匹配不到相似的问题，则将问题转到等待教师回答状态，并返回用户登录后的首页，提示用户：“无法回答，等待教师进行答复！”；

6) 学生可以对答疑的结果进行评价，满意或者不满意。满意问题解决，并返回用户登录首页；

7) 若用户对智能答疑给出的答案不满意，则系统将问题转到等待教师回答的状态。返回用户登录首页，提示用户：“无法回答，等待教师进行答复！”。

我要提问流程图如图 4-5 所示：

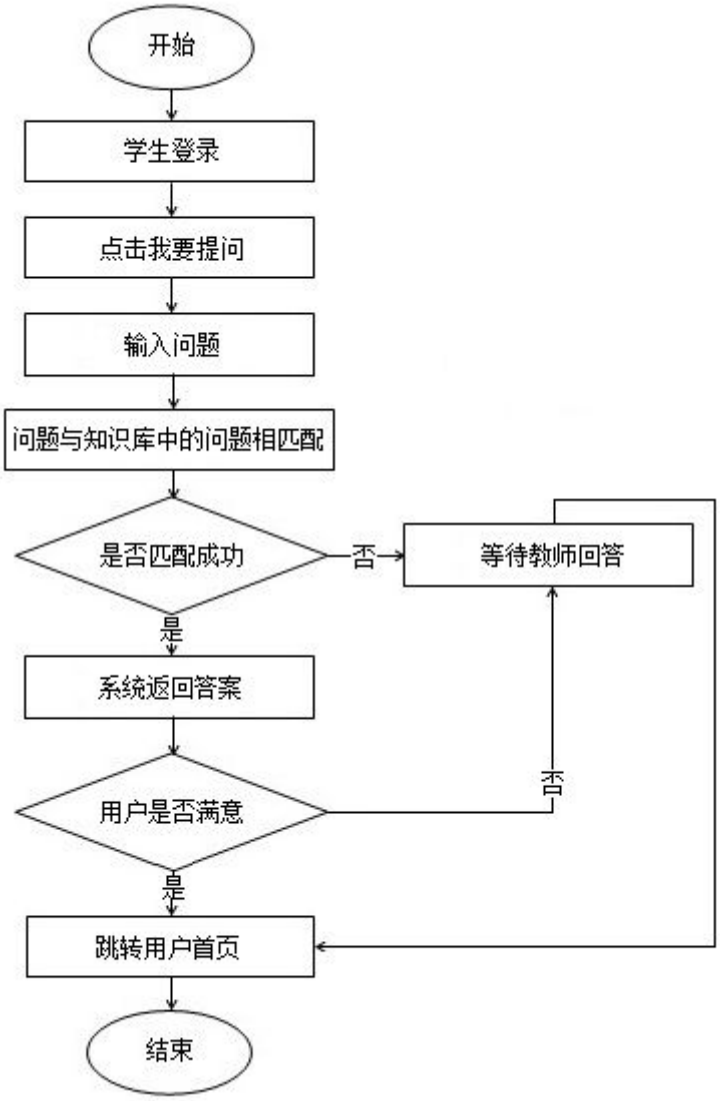

图 4-5　我要提问流程图

4.2.3　教师用户模块

教师用户模块主要包括问答列表和等待回答两个子功能模块。

(1)　等待回答　教师回答问题的时候，首先登录系统，然后在等待回答页面查看学生提交待回答的问题，之后教师对每个问题进行回答，回答完毕后，最后将问题和答案一并提交到知识库中。

主要流程如下：

1) 以教师身份登录后，在首页的页面左上角点击等待回答按钮；

2) 点击后，后台将请求连接到数据库的待回答表；将读取到的未被回答的问题反馈到前台，并跳转页面；

3) 教师选择待回答的问题，进入回答页面；

4) 教师需要根据待回答的问题输入准确的答案，随后提交答案，实现系统更新答案的命令；

5) 答疑系统在收到教师提交的答案更新后，将更新后的问题与对应答案一同存储进入数据库，并跳转回问答列表页面。

等待回答流程图如图 4-6 所示：

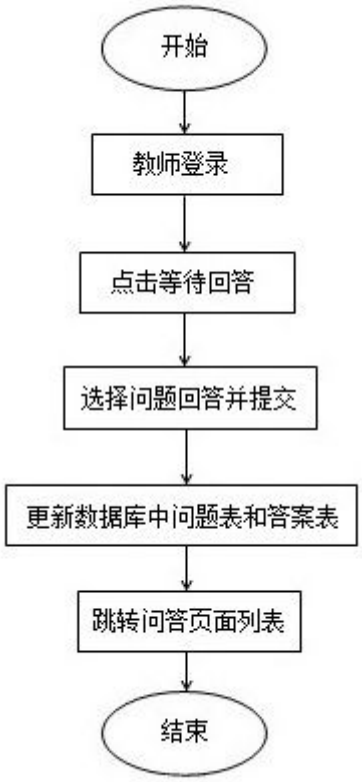

图 4-6　等待回答流程图

4.2.4　管理员用户模块

管理员用户模块主要是对账户进行管理，分为两类：用户管理和删除账户。用户管理可以更改用户密码及相关属性信息。对于不再使用的账户，管理员通过删除用户对其进行删除操作。

4.2.5　问答列表模块

在登录系统时，使用教师或者学生权限的用户登录，在首页左上角均能显示问答列表选项按钮。点击后，所有访问过系统的用户所提出的问题以及相应的答案均显示在页面上。

问答列表流程图如图 4-7 所示：

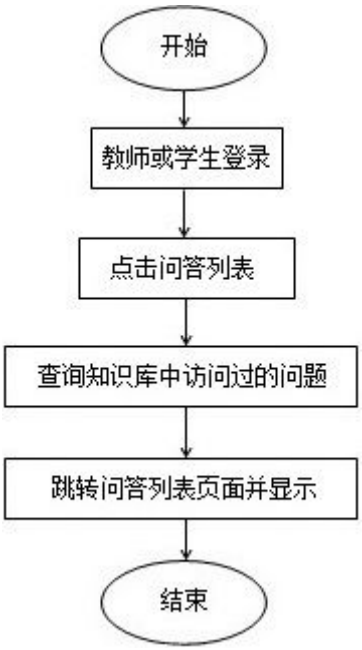

图 4-7　问答列表流程图

4.2.6　常见问题模块

常见问题就是系统知识库中的 FAQ[45]表里的问题。在登录系统后，进入登录页面点击问答列表按钮后，会在左上角显示常见问题按钮。点击常见问题按钮，所有常见问题的信息均显示在页面上。常见问题流程图如图 4-8 所示：

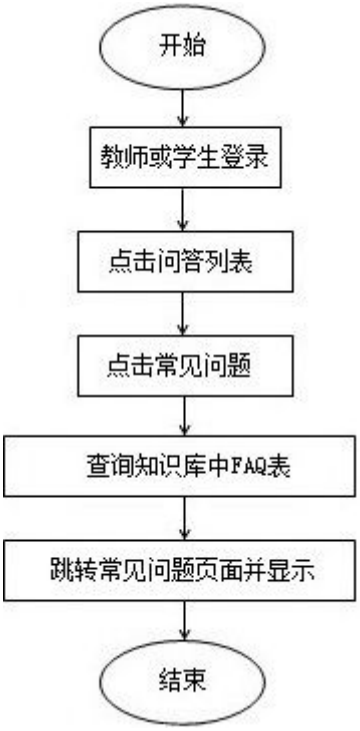

图 4-8　常见问题流程图

4.3　数据库设计

众所周知，系统开发的核心与基础就是一个强大的数据库[46]。数据库的设计是建立一个与其对应开发应用系统的向匹配的数据库，在整个信息开发中属于重中之重。数据库设计[47]，实现了重新整合开发信息系统中科用到的巨量的构架化数据，并在计算机中储存起来，数据库系统可以实现对整个信息的维护、储存以及使用过程的检索。信息系统利用相关的数据库软件，可以实现大数据的储存以及数据信息的管理，在需要提取相关信息时，可以及时准确的在庞大的数据信息中将其分离出来。不得不说，数据库设计是一个系统开发与运行的关键节点，相当于一个楼层的地基结构，如果数据库的设计出现差错，很可能造成数据库表的关联修改、索引的建立等一系列的问题，将会为系统开发以后的维护带来大量的工作。

数据库设计(Database Design)是指基于一定具体的数据库依据用户的使用要求实现这一数据库的表结构与建立全新数据库的全部流程。因为系统的整体需求需要逐渐完善，逐步进行，系统也要一步步进行一系列修改测试，故为了满足与数据库相关软件程序的运行实现，数据库设计内容也随之更正，所以数据库的最佳设计不会一蹴而就，这将是一个反复测试反复求证的过程。数据库的设计作为整个系统也研发的初始阶段，开始的每一步都要认真对待、查漏补缺。数据库设计主要基于如下两个主要的原则：

1) 数据库逻辑设计是对数据进行数学建模，建立数据模型的一个步骤，这个设计和数据库本身没有联系，只是单纯的体现需求运行的逻辑关系，包含了所有实际运用以及相互关系，属于具体业务与实体及关系之间的衔接对应。

2) 数据库物理设计是根据前部分逻辑设计决定了在物理设计步骤的两种选择，即关系数据库和面向对象数据库，并进一步确定与之相符的数据库表结构。在市场上关系数据库占绝大多数。

数据库逻辑设计是属于整个数据库设计的前期工作，主要囊括确定实体、实体属性和实体间关系等，以及确定设计需要的数据表内容的工作。数据库逻辑设计是强调了数据库以及数据库所应用实现的整体性能，其中相应数据库表的对应以及各实体之间的逻辑关系、表主键是重中之重。数据库设计的缺陷将直接导致数据库乃至整个系统的性能问题，也为后期扩大数据库内容以及使用过程的检索增加难度。良好的数据库中明确了实体与关系，为后期程序开发中代码的设计编排提供稳定的基础，也为后期数据库的索引调优提供了方便。

针对本文教学网站智能答疑系统中三种用户对系统的使用过程，我们来对系统中需要的数据表进行总结。

首先，系统中三类用户对系统的使用过程如图 4-9 所示：教师用户通过回答问题与系统中的答案表进行交互，学生用户通过输入问题，查询答案，对系统中的问题

表和答案表进行操作，另外，教师用户和学生用户还与系统中的用户表直接关联，管理员可以对该用户表进行操作。

图 4-9　系统中三类用户对系统的使用

根据分析系统功能需求，来实现对数据库表的设计的时候，考虑到以后系统数据库维护的方便性，数据库的关系表外的外键是必不可少的一部分，外键在一定程度上提高了表与表的相关性，也实现了后期维护方便快捷的目的。因为数据表的修改实现了牵一发而动全身，在修改一张表的同时，其他相关性表将自动实现修改。数据表间的关系图如图 4-10 所示：

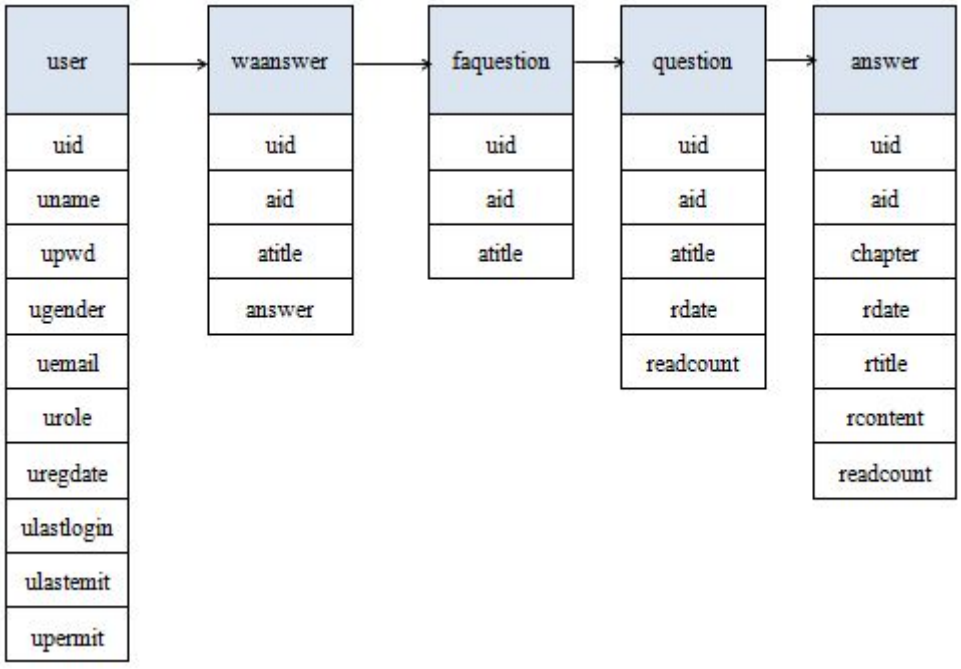

图 4-10　数据库关系图

主要的数据项和数据结构如下：

(1)　用户表(user)　主要用来管理用户登录类型，用户的描述和其所拥有的权限内容有：用户 id，用户名，密码，性别，邮箱，职称，注册时间等。

表 4-1　系统用户表结构图

列名	数据类型	允许空	说明
uid	可变字符	否	ID 标识
uname	可变字符	否	用户名
upwd	可变字符	否	密码
ugender	字符	否	性别
uemail	可变字符	否	邮箱
urole	整数	否	职称
uregdate	日期	否	注册时间
ulastlogin	日期	是	登录时间
ulastemit	日期	是	提交时间
upermit	整数	是	用户权限

(2)　问题表(question)　主要用来存储所有问题的基本信息，主要内容有：id 标识，问题 id，问题标题，访问次数等内容。

问题表是智能答疑系统中最关键的一个表。该表存储了针对某门课程的所有问题。问题表中存放的问题主要是由人工输入产生的，后期主要由教师和管理员进行问题表的维护。当学生在该系统中进行问题提问时，系统首先对该问题进行中文分词处理，其次进行语句相似度的计算和排序，得到相似度值最大的一个问题，最终通过此问题标识对应的答案表返回答案。

表 4-2　问题表结构图

列名	数据类型	允许空	说明
uid	可变字符	是	ID 标识
aid	整数	否	问题标识
atitle	可变字符	否	问题标题
rdate	日期	否	建立日期
readcount	整数	是	访问次数

(3)　常见问题表(faquestion)　主要存储常见的一些问题，包括 id 标识，问题标识和问题标题。常见问题表简称 FAQ 表，主要存储针对某门课程学生经常提出的一些问题汇总。在问题表的结构中我们有统计访问次数，所以针对此项信息我们把学生经常提问的问题进行汇总，输入到常见问题表中。后期主要由教师和管理员进行维护。它和问题表一起构成了系统的知识库。

表 4-3　常见问题表结构图

列名	数据类型	允许空	说明
uid	可变字符	是	ID 标识
aid	整数	否	问题标识
atitle	可变字符	否	问题标题

(4)　答案表(answer)　主要用于存储答案的相关信息，主要包括：问题 id，答案的内容，用户 id，所属章节，回答日期，对应的问题，访问次数等内容。

答案表主要用于存放知识库中问题的答案，每个答案与对应的问题关联，并记录了该答案生成的时间等信息。当学生在系统中输入问题进行问题咨询时，系统首先查找问题表，对问题进行中文分词处理，和语句相似度匹配。如果匹配到该问题，那么就在答案表中对该问题的答案进行查找，从而返回该问题对应的答案。

表 4-4　答案表结构图

列名	数据类型	允许空	说明
uid	可变字符	是	ID 标识
aid	整数	是	问题标识
chapter	整数	否	问题属于章节
rdate	日期	否	建立日期
rtitle	可变字符	是	问题标题
rcontent	可变字符	是	问题答案
readcount	整数	是	访问次数

(5)　待回答表(waanswer)　待回答表是临时存放问题的表。包括 id 标识，问题标识，问题标题和是否回答。

学生在登录系统后，在页面上输入问题内容，并点击提交；系统将问题请求转到后台，并在知识库中匹配相似问题；若系统匹配不到相似的问题，则将问题转到等待教师回答状态。此时，该问题被记录到待回答表中，等待回答。当教师进行答复后，点击提交，此问题标识已回答。

表 4-5　待回答表结构图

列名	数据类型	允许空	说明
uid	可变字符	是	ID 标识
aid	整数	否	问题标识
atitle	可变字符	否	问题标题
answer	字符	否	是否回答

4.4　本章小结

本章主要是对系统功能模块设计进行了详细说明，包含了用户的注册和登录模

块，教师、学生、管理员三类用户模块，以及常见问题和问答列表模块。另外根据三类用户的使用，我们对系统中需要的数据库进行了设计，为后续章节的系统实现做了铺垫。

第 5 章　系统关键技术原理与实现

本章根据教学网站智能答疑系统的需求分析和设计情况，主要对教学网站智能答疑系统中各个子模块的功能[48]进行实现，包括以下几部分：用户管理、问题查询、问题匹配、问题显示、问题维护，如下图 5-1 所示。其中问题查询和问题匹配是该系统中最重要的部分，问题查询主要涉及对自然语言提出的问题进行中文分词的预处理，然后根据预处理得到的分词在问题表中进行问题查找。查找过程中涉及到一个关键的技术是问题匹配，这里采用语句相似度的匹配算法[49]进行问题匹配。

该部分以工作流模型为基础，本文中的智能答疑系统基于 B/S 架构模式下，采用 Java EE 技术，Apache tomcat 作为网站服务器，SQL Server 2012 作为后台数据库。在尽可能低成本、高效率情况下，能够满足教学网站智能答疑的业务需求。

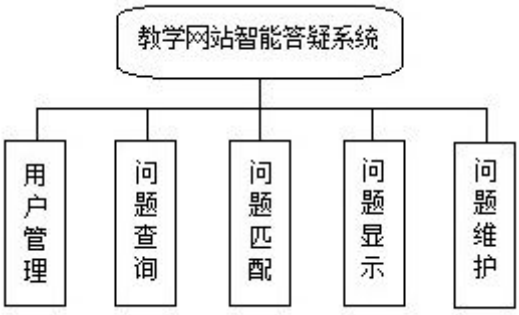

图 5-1　系统实现包括的功能模块

5.1　中文分词处理算法

在系统中输入需要检索的汉语问题之后，答疑系统能够实现对输入的汉语进行汉语的中文分词。中文分词处理算法的方式并不复杂，就是将输入的中文句子按照句子成分拆分成多个独立的汉语词语，也可称之为特征词。

5.1.1　中文分词算法的基本原则

(1)　粒度最大原则　所谓粒度最大，即在对一段中文内容，或一个中文语句进行分词的过程中，而每个中文词包含的中文字符数越多越好。这样，粒度越大，所表达的含义越精确，进而减少歧义的产生。如:“交通局长”可以有三种拆分方式：1、交通局长；2、交通和局长；3 交通局和长。系统会根据语意的意境，在系统所使用的词库的众多特征词中拆分出最好的分词效果。

(2)　分词最少原则　即总体词数越少越好。在相同字数的情况下，总词数越少，单个词组的粒度越大，分解出的语义单元越少，相对应的单个语义单元的权重就会越大，因此产生的对语义理解的歧义越小，准确性会越高。

(3) 最大包含原则 在切分输入的相关汉语句子时，切分出的词语中应当含有的非词典词语数量应当尽量最小，非词典词－词典中没有收录的词语与单字。切分的结果应当含有尽量多的单字典词，主要为可以独立使用的一些单字，如“得”、“与”、“是”、“你”、“们”、“她”等。

5.1.2 中文分词算法

根据现有技术，依照基于内容不同较为普遍的分词算法[50]主要有 4 种：①基于字符串匹配的分词方法；②基于理解的分词方法；③基于统计的分词方法；④基于规则的分词方法。

在此答疑系统中主要使用基于字符串匹配的分词方法来对用户输入中文问题首先进行中文分词的预处理，便于在知识库中对该问题进行匹配和查找，进而进行对应问题答案的返回。

基于字符串匹配的分词方法主要就是解析输入的中文串字符，然后对字符串内容与已有的数据库中词条进相似对比，假如在数据库中找到该词条字符，即表示字符串匹配成功。字符串匹配的匹配方式也有很多种，基于扫描方式分为正、逆向匹配方式，如果匹配按照字符串的内容多少来进行匹配，则分为最长、短匹配；按照是否与词性标注过程相结合，又可以分为单纯分词方法和分词与标注相结合的一体化方法。在应用中，最常见的主要有三种分词方法：正向（左—右）最大匹配法，逆向（右—左）最大匹配法以及将字符串中实现词数最小的最少切分法。考虑到逆向匹配采用逆向逐字匹配，这样词句不会产生误解的现象，从精准度上来说成为智能答疑系统设计的最优选择。

逆向最大匹配算法，按照从右向左的顺序，循环扫描字符串，同时与后台的关键词进行模式匹配。如匹配成功，则将该字符串作为关键词；若匹配失败，则依次减字，直到词典命中或剩下一个单字。

其算法流程如下：

1) 长句变短句。即通过标点符号，将中文内容或一个长句，先拆分为若干个短句（即最大词组）。

2) 读取短句或词组字符串长度S1，即短句或词组所包含的字数。

3) 设定阈值MaxLen，作为拆分截取得到的词组长度的上限。

4) 从短句或词组S1中，截取长度为字数减去阈值到字数的字符串W，在词典中检索该词时候存在。若存在，执行(5)；否则，执行(6)。

5) 记住该字符串记为S2，将长度从字数逐步减少到字数减去阈值，继续执行(4)，直到长度为“0”。

6) 阈值减1，再次执行(4)。

逆向最大匹配算法程序流程图如图 5-2 所示：

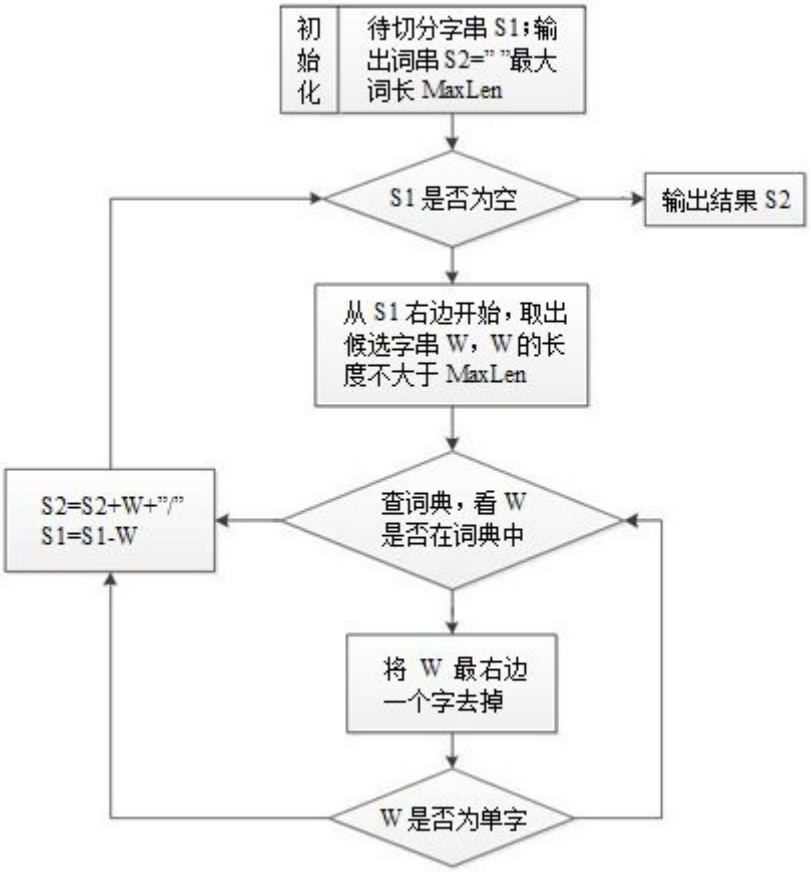

图 5-2　逆向最大匹配算法流程图

5.2　语句相似度计算

在用户输入自然语言描述的问题时，本系统首先需要根据系统中的知识库来与用户输入的问题进行匹配，将知识库中的与用户输入的问题对应的问题答案输出作为用户查询的问题的答案。因此，首先需要对用户输入的中文语句进行分词处理，从而来分析两个语句之间的相似性。语句相似度[51]是表示的是这两个句子的雷同程度。当系统判定这两个句子达到一定的相似度时，可以认定所选的两个句子为相似语句。句子的雷同字、长度、词形以及词语顺序都对相似度判定有着影响。

在对问题进行相似度匹配之前，我们首先提出计算语句相似度的几个指标，然后给出语句相似度匹配算法的执行过程。

5.2.1　语句相似度计算指标

(1)　词形相似度　用户输入的自然语言描述的问题L(文中以汉字为例)中的一个语句S(Sentence)是L中的单字和特殊符号(以下简称单字)的一个有序集合。该语句S的长度即是S中单字的个数，此处用Len(*S*)表示。SameWC(*X*，*Y*)表示用户输入的语句和知识库中的问题的语句中相同单字的个数，只有当两个语句中出现的单字的数目越多时，则说明这两个语句的相似度越高。在一个自然语言描述的语句中，通常

只有语句中的单字才有具体的意义，而特殊符号通常只是表示该语句的感情色彩，一般不会影响该语句的含义，因此在这里我们仅仅从用户输入的自然语言描述的问题和知识库中问题描述之间的共同的单字来描述两个语句的相似度。

比如以下两个问句：“什么是栈”与“什么是列队”。“什么是栈”这条语句可以划分为“什么”，“是”，“栈”这三个单字；对于第二条语句来说，可以划分为“什么”，“是”，“列队”这三个单字，那么单单从词形的相似度来说这两条语句的前两个单字都是相同的，因此具有一定的相似度。

针对一般的情况，对于两个自然语言的句子来说，在计算词形相似度时，首先要对自然语言进行词语切割，这些词语主要包括名词，动词，形容词等，其中名词和动词作为主要的单字进行相似度匹配[52]。这是因为在一条语句中，名词和动词表示这条语句的主要含义，而其他的形容词，介词等都只是起到修饰的作用，因为为了提高匹配的效率，我们提出主要对用户输入问题的名词和动词进行词语切割，然后对知识库中的问题也进行相应的词语切割，然后进行单词匹配。在两个待匹配的两条语句中，当一个单字在X，Y中出现的次数不同时，以出现次数少的计数。语句X，Y的词形相似度WordSim(X，Y)由公式(5-1)决定：

$$WordSim(X，Y)=SameWC(X,Y)/Max(Len(X),Len(Y)) \quad (5\text{-}1)$$

意义：上述公式表示了在用户输入的自然语言的问题中，如果输入的问题与知识库中的语句中相同的字数越多，则我们认为这两个语句越相似。

(2) 语句长度相似度 在中文的自然语言相似度计算中，两条语句之间的语句长度也能从一定程度上来说明这两条语句之间的相似度。比如，“我是中国人”与“我们中国人是勤劳勇敢的”，那么单单从语句长度的相似度上来看，这两条语句就是完全不同的两句话，因为我们这里从用户输入的自然语言的语句长度与知识库中问题的语句长度进行比对，从直观上来初步判断两条语句之间的相似度。需要注意的是，我们在这里从语句长度进行的相似度判断，仅仅作为一个辅助的手段进行的语句之间的相似度判断，这是因为绝大多数语句长度相似的两条语句，其中表达的语句的意思是不同的，比如中国古诗中每句诗词的长度都是固定的，但是所表达的意义都是不一样的，但是我们在这里主要从反向角度来判断，即用户输入的自然语言的语句长度与知识库中问题的语句长度相同时，我们不能说这两句话是相似的，但是相反的，如果用户输入的自然语言的语句长度与知识库中问题的语句长度不一致时，则我们认为这两句话在很大程度上是不一致的。

在这里我们采用Len(X)，Len(Y)分别表示用户输入的查询语句X和知识库中每个知识对应的问题语句Y的长度，即两个语句中的单字的个数。在这里，我们把语句长度相似度LenSim(X，Y)由公式(5-2)决定：

$$LenSim(X，Y)=1- abs(Len(X)-Len(Y))/(Len(X)+Len(Y)) \quad (5\text{-}2)$$

意义：上述公式直接从两个语句的长度进行相似度判断，即两个语句的长度越接近，两个语句越相似。需要说明的是这里所说的问题的长度不是指每个语句中包含的字数，而是从单字的角度首先对用户输入的自然语言和知识库中的自然语言进行词语切割，然后根据词语的个数来判断语句的长度。

(3)　词序的相似度　在中文自然语言的语句中，通常两个词语之间的顺序也会对这两个语句之间的相似度产生影响。这是因为中文语言与英文语言不同的地方，对于英文的语句来说，通常主动句和被动句都会大量使用，而对于中文来说，通常都使用主动状态，那么对于用户输入的自然语言的语句与知识库中问题的语句如果有些词语的顺序存在不一致的话，那么在很大程度上将说明这两句话表达的不是同一个意思，因为也就说明知识库中该问题描述对应的知识不是用户输入的查询问题的答案。

在这里，我们通过公式来定义这种关系，我们使用Onews(*X*，*Y*)表示在X，Y中都出现且都只出现一次的单字的集合 Pone(*X*，*Y*)表示Onews(*X*，*Y*)中的单字在x中的位置序号构成的向量，Ptwo(*X*，*Y*)表示Pone(*X*，*Y*)中的分量按对应单词在Y中的次序排列生成的向量。

语句 X，Y 的词序相似度由公式(5-3)决定：

$$Ordersim(X,Y)=\begin{cases}1-Reward(X,Y)/|Onews(X,Y)-1| & \text{当 } Onews(X,Y)>1 \text{ 时}\\ 1 & \text{当 } Onews(X,Y)=1 \text{ 时}\\ 0 & \text{当 } Onews(X,Y)=0 \text{ 时}\end{cases} \quad (5\text{-}3)$$

这样定义词序相似度的优点是：当一个分句或短语整体发生长距离移动后．仍与原来的语句很相似。实现快捷，算法复杂度为O(*m*)，其中m=Onews(*X*，*Y*)。另外需要说明的是，我们这里采用的词序的相似度也需要提前对用户输入的自然语言的语句与知识库中问题的语句进行单字划分，然后再对每个涉及到的词语进行词序相似度计算。

5.2.2　语句相似度计算

基于上一节中提出的语句相似度计算指标，下面介绍问题匹配的流程。两个语句之间特征词或关键词的相似度越高，即词形相似度（计算算法，参照 5.1.2 中文分词算法）（WS=K1WordSim(*X*,*Y*)）越高 ，WS 越趋近于 1；两个语句之间，长度越接近则其语句长度相似度（LS=K2LenSim(*X*,*Y*)）越高，LS 越趋近于 1；同时根据所提问题语句的特征词顺序，与知识库中已有语句的特征词顺序进行对比并计算，得到其语序相似度（OS=K3OrderSim(*X*,*Y*)），OS 越趋近于 1。两条语句之间的语句相似度（SS=WS+LS+OS）。其中 K1，K2，K3 是常数，且满足 K1+K2 +K3 =1，显然 SS∈[0,l]。

在语句相似度中我们能够理解词形相似度起着主要作用，语句长度相似度和词

序相似度起着次要的作用，因此 K1，K2，K3 取值时应该有 K1>>K2>K3。在本系统中我们设置的值为 K1=0.7，K2=0.2，K3=0.1。

如下图 5-3 所示：

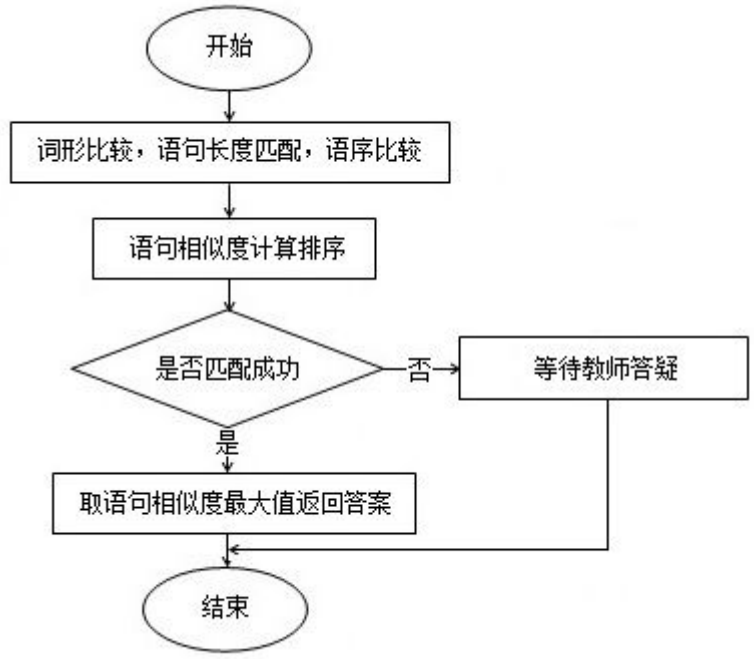

图 5-3　问题匹配流程图

在答疑系统中，用户输入一个问题后，首先在系统已有的知识库中根据关键字进行问题匹配。这里主要从语句长度、语序以及语句相似度三个方面进行匹配和比较，并记录结果。当完成对知识库的检索时，对其中的语句相似度权值按降序进行排序。如果第一条语句的相似度权值大于或等于预设的语句相似度阈值时，则认为两条语句语义相同，并输出该问题对应的答案，作为本次智能答疑问题的答案；如果第一条语句的相似度权值小于预设的语句相似度阈值时，系统将该问题，作为待回答问题输出到教师的“等待回答”列表中，同时向学生用户提示：“无法回答，等待教师进行答复！”。（在本系统中，我们预设的语句相似度阈值为 0.5）

在这里，我们把系统中问题的匹配过程看作是学生用户输入问题和数据库中已有问题两个语句之间的相似度计算的问题。

设输入的问句为 Input，知识库中的问题集为 Q，q 是知识库中的一个问句，$q \in Q$，则问题匹配的过程可以用公式(5-4)描述：

$$Q'=\text{argmaxSim}(\text{1nput}, q)(q \in Q) \tag{5-4}$$

其中：Q’表示找到知识库中与输入问题最相似的问句。根据以上公式，要查找与 Input 最相似的语句，需计算 Input 与知识库中的所有问句的相似度，从中选出最大的一个。如果采用遍历法，许多与 Input 相似度为 0 或相似度很低的都要参与计算，算法效率低下，而且会受到知识库规模的影响。为此，本文提出在常见问题集中建

立候选问题集和基于单字的倒排索引表和语句长度表来解决知识库中的匹配速度问题。

建立候选问题集的目的是缩小查找范围，使后续的相似度计算在相对较小范围内进行。从以上相似度计算公式中可以看出，与 Input 相似度最大的语句的 SameWC(Input，q)值也就较大，所以从所有的知识库问题中选出与问题相同的字数的前 k1(通常设为知识库问题总数的 50%)个就会包含与 q 最相似的语句。我们把这前 k1 个问题组成候选问题集。候选问题集的建立本质上就是一个求 SameWC(Input，q)的最大值的前 k1 个的集合。

在候选问题集中计算前 k1 个 SameWC(Input，q)时，如果将知识库中的问句一一读出来和 Input 进行比较，效率比较低。为了能够快速统计知识库中究竟有多少问句含有某个字，设计了如下数据结构如图 5-4 所示：

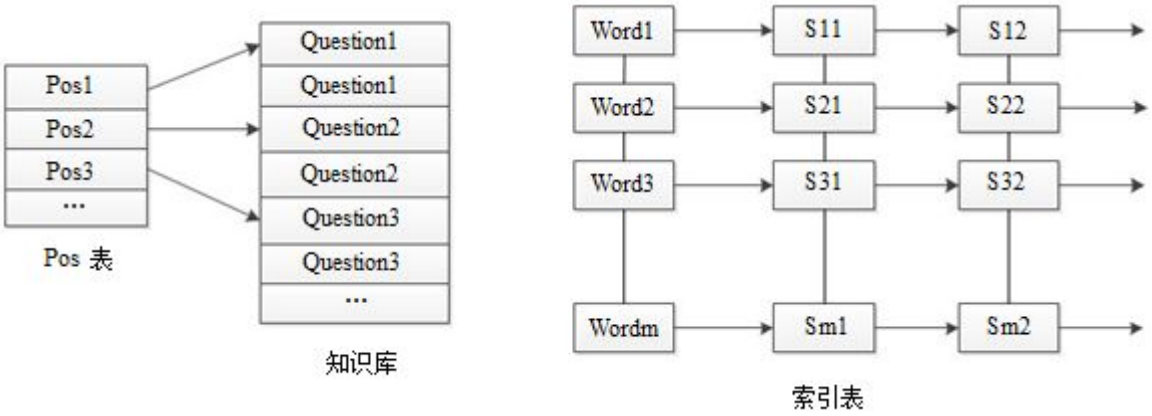

图 5-4　候选问题索引表

图中Pos表记录了知识库中每个问句在库文件中的位置，索引表中的Word1，Word2．Word3，... ，Wordm是知识库中的问句所包含的词经过排序后所形成的链表，每个Wordi指向一个s链表，这个s链表中的每个节点记录知识库中含有Wordi的一个问句的语句号。

算法1：单字倒排索引表和语句长度表建立算法。

输入：Q

输出：单字倒排索引表InvTab和语句长度表LenTab

1) For each q in Q；

2) 求出q的偏移量pos(q)；

3) 把pos(q)及Len(q)插入LenTab表中；

4) for each word in q；

5) 把pos(q)插入到InvTablhashW(word)l~lLINKword中；

6) 输出InvTab和LenTab。

算法中的Q为知识库中所有问句的集合，可以用数组结构来表示。InvTab为单字倒排索引表，其每个记录为二元组<word，ILINKWord>，其中ILINKWord为单字word出现的语句偏移量的链表，该索引表按单字以散列方式组织，能够实现快速建表和查找。设散列函数为hashW(word)，可用任何一种方法解决散列地址冲突问题。为使算法描术简明，这里假设没有冲突，即可直接通过lnvTab[hashW(word)J-qLINKword存取word的索引集合。语句长度表LenTab的每个记录为二元组<pos，fen>，其中pos表示一个语句的偏移量，Ien表示该语句的长度。该表也以散列方式组织。设散列函数hashpos(pos)，即可通过LenTab[hashpos(pos)]->len存取偏移量为pos的语句的长度。单字倒排索引表作用是在不读知识库的情况下就可计算SameWC (Input，q)，而且与Input相似度为0的语句不参与计算。语句长度表的作用是在不读知识库的情况下就可在已算出从大到小依次的前ki 个SameWC (Input，q)(即在候选问题集中)的基础上计算WordSim(1nput，q)。

5.3 本章小结

本章主要对智能答疑系统用到的关键技术进行了详细的阐述。包括中文分词技术所采用的算法，语句相似度计算的三项指标，以及根据计算公式得出语句相似度值。本章节涉及到的两种算法作为系统核心起到重要的作用。

第 6 章　智能答疑系统实现

6.1　开发工具与环境

智能答疑系统需具备如下硬件和软件环境需求：

硬件需求：CPU 主频至少单核 1Ghz 以上，内存至少 512M 以上，硬盘空间至少 30G 以上。

软件需求：

1) 操作系统：Windows 7

2) 系统开发平台：采用 Java EE 技术，开发工具使用 My Eclipse，Apache tomcat 作为网站服务器。

3) 数据库服务器：Microsoft SQL Server 2012

4) IE 浏览器：IE、360、搜狗等其他浏览器

本文的智能答疑系统物理架构如图 6-1 所示：

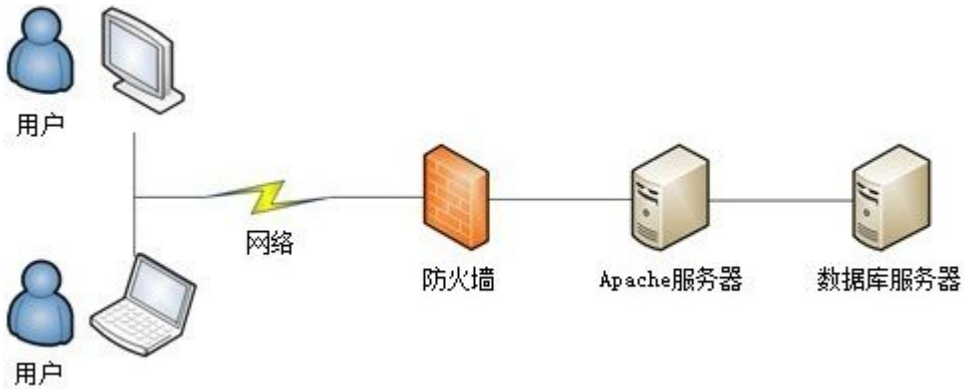

图 6-1　系统物理架构图

如图 6-1 所示，客户端借助互联网在对答疑系统的两个服务器数据进行访问时，首先要通过防护墙的审核。防火墙起着将外部互联网与内部局域网相互隔离的作用，以防外部网络的病毒侵入，篡改数据库内容。对于校园内部网络用户可直接访问服务器；如外部网络用户需要访问时，需要打开服务器的相应端口，以便外网对系统的访问。Apache 服务器通过访问数据库服务器，与其进行数据交互。

6.2　系统功能实现

本系统数据来源是人民邮电出版社出版的《数据结构》一书，出版时间是 2016 年 1 月，罗福强、杨剑等编著。根据第四章智能答疑系统总体设计，本章分为以下六个部分来实现系统的功能：系统登录、问题查询、等待回答、问答列表、常见问题和我的问题。

6.2.1 系统登录

系统登录页面如图 6-2 所示：

图 6-2 系统登录页面

用户登录系统的时候，在登录界面输入用户名和密码，选择登录类型，点击“登录”即可。在使用者第一登录该系统时，首先要进行身份的注册，选择“请从这里注册”按钮。注册过程的详细信息都需要填写，以便系统核实下次登录后，使用系统相关功能的权限。

6.2.2 问题查询

图 6-3 学生查询页面

如上图所示，系统中以学生权限的用户登录后，首页上我们能看到我要提问按

钮。点击后，即显示提问的主题和内容输入框，学生根据自己的问题在搜索框中输入要查询的问题。

对于学生用户在登录自己主页面后，进行的问题查询的主要核心代码如下图所示：

```
// 构造查询对象
SolrQuery query = new SolrQuery();
// 设置查询字符串
if (condition instanceof CommonDocumentQuery) {// 通用查询
    CommonDocumentQuery commonCondition = (CommonDocumentQuery) condition;
    queryString = buildQueryString(buildCommonCondition(commonCondition));
    if (StringUtils.isNotBlank(commonCondition.getFilterString())) {
        query.setFilterQueries(commonCondition.getFilterString());
    }
} else if (condition instanceof AdvanceDocumentQuery) {// 高级查询
    AdvanceDocumentQuery advanceCondition = (AdvanceDocumentQuery) condition;
    queryString = buildQueryString(advanceCondition.getQueryString());
} else {
    return page;
}
query.setQuery(queryString);
query.add(CommonParams.FL, FL);

// 设置排序字段
for (Object obj : condition.getSortInfos()) {
    SortInfo sortInfo = (SortInfo) obj;
    query.addSortField(sortInfo.getColumnName(), ORDER.valueOf(sortInfo.getSortOrder()));
}

// 设置高亮参数
query.setHighlight(true);
query.setParam("hl.fl", Constant.SOLRJ_HL_FIELD);
query.setHighlightSimplePre(Constant.SOLRJ_HL_PRE);
query.setHighlightSimplePost(Constant.SOLRJ_HL_POST);
```

图 6-4　问题查询模块的核心代码

当输入完毕后点击提交，系统自动输出结果，如图 6-5 所示：

图 6-5　智能答疑页面

最后我们设置了满意或不满意两个选项，可以让学生对本次的答疑进行评价。我们可以作为参考方便系统管理员后期对问题库的维护和管理。最后结束本次提问。

6.2.3 等待回答

在教师用户登录到自己的主页面后，将会在自己的主页面上查看到学生提交上来的，在系统中无法自动解答的问题，这里称之为等待回答的问题，这些待解答的问题将根据学生提出问题的时间先后顺序进行排列。对于这些待解答的问题，教师可以从问题列表中任选一个进行回答。对于已经回答的问题包括回答，都一并提交到数据库，这样待解答问题里就不显示出该问题。当学生用户下次再进行问题查询的时候，系统直接从数据库中进行查找，从而更完善了系统本身的智能答疑的能力。

如图 6-6 所示以教师权限的用户登录后，在首页能显示等待回答按钮。点击后，即显示该问题的信息。

图 6-6　等待教师回答页面

教师点击回答后，即在该页面下方显示回答内容输入框。教师输入答案后并提交，如图 6-7 所示：

图 6-7　教师回答页面

经教师进行答复后，等待回答问题结果即显示在问答列表中，如图 6-8 所示：

图 6-8　教师答复结果页面

6.2.4　问答列表

我们在登录系统时，使用教师或者学生权限的用户登录，在首页左上角均能显示问答列表选项按钮。点击后，所有访问过系统的用户所提出的问题以及相应的答案均显示在页面上。如图 6-9 所示：

图 6-9　问题列表显示页面

6.2.5　常见问题

我们在登录系统，进入登录页面点击问答列表按钮后，会在左上角显示常见问题按钮。点击常见问题按钮，所有常见问题的相关信息均显示在页面上。如图 6-10

所示：

图 6-10　常见问题显示页面

6.2.6　我的问题

系统中以学生权限的用户登录后，首页上显示我的问题按钮。点击我的问题按钮后，所有该学生提问过的问题相关信息均显示在页面上。如图 6-11 所示：

图 6-11　我的问题显示页面

6.3　系统测试

6.3.1　测试目的

在整个系统开发的全部过程中，想要保证开发全部系统过程的稳定性与可靠程

度，在开发中的分析、设计与编码每一步都使用了很多技术手段。但是由于软件开发的产品是虚拟的，考虑到其是极度复杂、设计学科较多的一种集成产品，可能设计过程会存在部分错误。故在产品投入完成之前，完整的测试必不可少，软系统测试的主要内容有性能、功能、可用性、兼容性以及系统的安全性测试等环节。在测试完毕后，出具可靠的测试结果，研发者需要对测试结果进行汇总分析，解决其中发现的问题。

6.3.2　测试环境

服务器端：Windows 7、Apache tomcat、SQL Server 2012

客户端：Windows 7、IE 10

6.3.3　测试过程

(1)　基本测试　用于测试系统基本功能的实现情况和系统是否存在设计错误。基本测试过程：用户的注册及登陆，管理员对用户帐户的更新，删除操作，学生用户的查看问题答案功能以及查询问题，对不满意答案的提交等各类操作。教师用户对问题的查看及对某个问题的回答，提交到数据库等各类操作。检测页面生成情况及数据库连接情况。

(2)　容错性测试　用于测试系统对于用户在使用过程中对于填入的错误信息与违反越权相关法律法规指令的鉴别情况。测试过程如下：选择同一台计算机在智能系统登录时选择拥有不同权限的账号，在系统中复制如 URL 等非法请求页面，查看系统反应。然后可以使用不同的用户系统进行登录尝试，频繁进行非法操作、越权操作等，以进步一检验开发的系统对于非正常操作的处理能力。

(3)　并发性测试　主要用于在多用户在同时登录系统时，系统对于各用户操作产生的冲突处理能力。测试过程如下：在同的电脑上分别以不同的用户权限登录，此系统主要为学生、教师、管理人员、系统维护人员等等，然后同时进行系统的提交表单操作或者同步更改数据信息。按照规定进行系统功能、性能的操作，测试整个系统对于多事件的处理能力。

6.3.4　测试结果

经过对智能答疑系统的测试，系统完全正常运行，在众多分辨率的操作方式下，网页能够迅速打开并提供准确的网页地址。整个系统是所有基本功能可以完全实现，未出现乱码、闪退等错误情况。

基本测试，登录用户的功能全部实现，符合开发应用的需求。

容错性测试，系统对非法请求进行限制，对非法操作进行正确提示，限制非法用户访问页面。

并发性测试，多用户同时登录时未出现不正常状态，服务器对不同用户请求进行分步处理。

6.4 本章小结

本章介绍了智能答疑系统的硬件和软件环境需求，包含开发工具，运行环境和物理结构。对系统各个功能模块进行了展示，并进行了系统测试，运行正常。智能答疑系统的搭建促进了学生与教师之间的互动，并且学生也巩固了课程的知识，更进一步的提升了教学质量。

结 论

教学网站的智能答疑系统是学校信息化教学管理系统的组成部分之一，它结合学校开设的课程，利用内部网络平台，以 Web 界面形式向学生提供课程答疑功能。对于这个智能答疑系统要求，我们的目的是在满足用户最基本的答疑功能的同时，使学习者能够在教师不参与的前提下，利用智能答疑系统迅速准确的得到知识点的解答，从而最大程度的利用网络提升自学效率。同时，智能答疑系统直接替代了传统的教师人工对某一单一问题反复解答，极大的减轻了在线教师劳动强度。除此之外，智能答疑系统利用了大数据时代特点，对学生提出的问题加以归纳，重点、难点以数据化形式加以体现，在一定程度上提高了教师的网络教学质量。

本文论述了课程教学网站中智能答疑系统的重要性，对比了国内外具有代表性的智能答疑系统，参考已有的应用系统结构，提出了课程教学网站智能答疑系统的研究方案。分析了其技术难点，结合中文分词算法和语句相似度算法，构建起基于数据结构课程的智能答疑系统。在中文分词方面，采用逆向最大匹配算法实现问句分词的预处理；在语句相似度计算方面，主要从词型、语句长度和词序三个方面进行语句相似度的匹配和比较。最后结合软件开发技术将这两种算法应用到系统中。本系统采用了 B/S 架构模式，基于 Java EE 技术，使用 My Eclipse 作为开发工具，Apache tomcat 作为网站服务器，SQL Server 2012 作为后台数据库，构建起了智能答疑系统。满足用户对答疑的需求，并且在答疑测试中具有较好的运行效果。

本系统可以应用于专业化课程教学网站方面问题的智能答疑，虽然有比较不错的智能性，但是本系统仍然存在许多问题需要改进，对智能答疑系统的后续研究主要是以下几个方面：

1) 本系统研究了中文分词技术和语句相似度算法，但在算法细节上仍存一些问题，例如：在处理近义词或是存在歧义的词语时仍然存在一定的困难，而这就会导致其算法性能下降，所以有对算法进行进一步优化的空间。

2) 本系统是检索知识库中的问题来对学生进行答疑的，知识库的问题数量是关键。本系统为专业课程提供答疑尚且可以满足，但遇到海量的数据查询时数据库的设计和查询算法的实现就会需要相应的优化，否则在大数据环境下查询知识库中的数据就会成为整个系统的瓶颈。

3) 本系统在知识库中无法匹配搜索到的问题，提供了人工答疑的方法来解答；而当今全球互联网发展之迅猛，如何进一步利互联网资源来进行智能答疑，也是系统提升性能的研究方向。

目前对于答疑系统来说，其智能程度还无法达到人类一样能回答用户提出的各

种问题，还不具备人类的思维能力。所以，一旦知识库中缺少与问题匹配的相关内容，那么该系统就无法对用户提出的疑问给出满意的答案。但是智能答疑技术在近些年得到了很大的发展，许多技术日益趋于成熟，具有一定人工智能的答疑系统问世。在这些技术广阔的应用前景下，智能答疑技术也在不断向前快速发展，我们有理由相信在不久的将来对于智能答疑技术的研究必将会取得更大的进步。

参考文献

[1] 张伟远. 网络教学发展模式的理论构建与应用. 现代远程教育研究, 2013, (1): 7-14

[2] 彭婧. 一种用户交互的智能答疑系统的设计与实现. [湖南大学教育学硕士学位论文]. 长沙: 湖南大学, 2013

[3] 郭文俭. 基于课程教学网站的智能答疑系统的设计与实现. [吉林大学工程硕士学位论文]. 长春: 吉林大学, 2015

[4] 王萌, 俞士汶, 朱学锋. 自然语言处理技术及其教育应用. 数学的实践与认识, 2015 (20): 151-156

[5] 索俊锋, 刘勇, 邹松兵. 基于地理本体的综合语义相似度算法. 兰州大学学报, 2017, (53): 10-12

[6] 肖坤峨, 虞泉. 基于 WEB 的智能答疑系统的研究与构建. 软件, 2015 (6): 31-36

[7] E. J. EMANUEL. Online education: MOOCs taken by educated few. Nature, 2013, 503(7476): 342-342

[8] Y. C. KUO, A. E. WALKER, B. R. BELLAND, et al. A predictive study of student satisfaction in online education programs. The International Review of Research in Open and Distributed Learning, 2013, 14(1): 16-39

[9] S. PAL, A. CHAKRAVORTY. A Framework for Automatic Generation Of Answers to Conceptual Questions in Frequently Asked Question (FAQ) Based Question Answering System. International Journal of Advanced Research in Artificial Intelligence, 2012(1): 19-24

[10] 林桂亚. 基于 WAP 的课程教学网站设计与探索. 软件, 2013 (2): 26-27

[11] J. M. MULLER. Bus Rapid Transit: The Answer to Transport Problems in. Megacities: Our Global Urban Future, 2013: 179

[12] I. ELHALWANY, A. MOHAMMED, K. T. WASSIF, et al. Using Textual Case-based Reasoning in Intelligent Fatawa QA System. Int. Arab J. Inf. Technol., 2015, 12(5): 503-509

[13] M. C. I. NWOGUGU. Strategic Decisions, Risk Management and Strategic Alliances: The Case of Corporate Governance at Akamai Technologies; Openwave Systems; Novell; Ask. Com (Ask Jeeves); and Firstwave Technologies, Inc. 2015, 13(5): 204-210

[14] T. HAO, X. QIU, S. JIANG. Leveraging Semantic Labeling for Question Matching to Facilitate Question-Answer Archive Reuse//International Conference on Intelligent Computing. Springer, Cham, 2015: 65-75

[15] 张妍, 许云峰, 张立全. 基于云计算的中文分词研究. 河北科技大学学报, 2012, 33(03): 266-269

[16] 秦赞. 中文分词算法的研究与实现. [吉林大学工学硕士学位论文]. 长春: 吉林大学, 2016

[17] 刘欣. 智能答疑系统中句子相似度计算的研究与应用. [电子科技大学工学硕士学位论文]. 成都: 电子科技大学, 2011

[18] M. D. HANSON. The Client/Server Architecture. Server Management, 2000: 3

[19] 许锦. 基于本体的智能答疑系统研究与实现. [江西师范大学工学硕士学位论文]. 南昌: 江西师范大学, 2010

[20] 李宏波. 综合字典和统计分析的中文分词系统的研究与实现. [武汉理工大学工学硕士学位论文]. 武汉: 武汉理工大学, 2010

[21] 林冬盛. 中文分词算法的研究与实现. [西北大学理学硕士学位论文]. 西安: 西北大学, 2011

[22] G. H. FU, C. Y. KIT, J. J. WEBSTER. Chinese word segmentation as morpheme-based lexical chucking. Information Sciences, 2008, 178(9): 2282-2296

[23] 于重重, 操镭, 尹蔚彬, 等. 吕苏语口语标注语料的自动分词方法研究. 计算机应用研究, 2017, 34(5): 1325-1328

[24] 韩冬煦, 常宝宝. 中文分词模型的领域适应性方法. 计算机学报, 2015, 38(2): 272-281

[25] X. ZHANG, X. TANG, Y. ZHAO, M. LI. Research in intelligent search on Internet based on ontology. Computer Engineering and Design, 2006, 27(07): 1194-1210

[26] 李琳. 基于本体的 Multi-Agent 智能答疑系统的设计与实现. [宁夏大学工学硕士学位论文]. 银川: 宁夏大学, 2011

[27] 岳群琴. 计算机基础课程的智能答疑系统的研究与设计. [西南交通大学教育学硕士学位论文]. 成都: 西南交通大学, 2014

[28] 王保民, 刘明生, 邢飞. 基于语义的语句相似度计算研究. 河北科技大学学报, 2011, 32(04): 364-367

[29] 田久乐, 赵蔚. 基于同义词林的词语相似度计算方法. 吉林大学学报, 2010, 11

[30] 朱新华, 马润聪, 孙柳, 等. 基于知网与词林的词语语义相似度计算. 中文信息学报, 2016, 30(4): 29-36

[31] 田家旗. Java 开发语言的开发平台与 J2EE 编程技术问题研究. 信息技术与信息化, 2016, (4): 112-113

[32] 郝玉龙, 周旋. JavaEE 核心技术与应用. 第 3 版. 北京: 电子工业出版社, 2015:16-136

[33] B. KURNIAWAN, P. DECK. 深入剖析 Tomcat. 北京: 机械工业出版社, 2011: 22-189

[34] C. DIACONU, C. FREEDMAN, E. ISMERT, et al. Hekaton: SQL server's memory-optimized OLTP engine//Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data. ACM, 2013: 1243-1254

[35] A. ABELLO, O. ROMERO, T. B. PEDERSEN, et al. Using semantic web technologies for

exploratory OLAP: a survey. IEEE transactions on knowledge and data engineering, 2015, 27(2): 571-588

[36] 李超, 谢坤武. 软件需求分析方法研究进展. 湖北民族学院学报, 2013, 31(2): 204-211

[37] L. CHUAN. Improved Intelligent Answering System Research and Design. Advances in Intelligent and Soft Computing, 2012, 11(6): 583-589

[38] 石彦, 桂志海. 软件工程需求分析与质量保障. 无线互联科技, 2014, (2): 45-45

[39] 杜丽萍, 李晓戈, 于根, 等. 基于互信息改进算法的新词发现对中文分词系统改进. 北京大学学报(自然科学版), 2016, 52(1): 35-40

[40] 张立岩, 张世民. 基于语义相似度的主观题评分算法研究. 河北科技大学学报, 2012, 33(03): 263-265

[41] J. WHITTLE, J. HUTCHINSON, M. ROUNCEFIELD. The state of practice in model-driven engineering. IEEE software, 2014, 31(3): 79-85

[42] T. GUNAWARDENA, N. PATHIRANA, M. LOKUHETTI, et al. Performance Evaluation Techniques for an Automatic Question Answering System. International Journal of Machine Learning and Computing, 2015, 5(4): 294

[43] 刘兴. 基于 web 的智能答疑系统的研究与设计. [电子科技大学工程硕士学位论文]. 成都: 电子科技大学, 2013

[44] 刘松平. 智能答疑平台的研究与实现. [湖南大学工程硕士学位论文]. 长沙: 湖南大学, 2012

[45] X. WANG, H. HAN. Chinese Question Answering System Based on the Chinese FAQ[C]//2014 International Conference on Information, Business and Education Technology (ICIBET 2014). Atlantis Press, 2014, 68-73

[46] 王卫红, 田忠和, 曹玉辉. 数据仓库技术的产生及发展. 河北工业科技, 2002, (02): 59-62

[47] T. U. O. CHAO. Techniques of SQL Server Database Field Connection and Simplification. Computer Programming Skills & Maintenance, 2013, 12: 014

[48] J. XU, Y. Q. Li. Design and Implementation of Intelligent Question Answering System Based on Ontology. 2010 Second International Conference on Computational Intelligence and Natural Computing (CINC 2010), 2010, 9

[49] 李旭锋. 中文问答系统中问句理解和相似度计算的研究与实现. [华南理工大学工学硕士学位论文]. 广州: 华南理工大学, 2010

[50] M. Y. ZHANG, Z. D. LU, C. Y. ZOU. A Chinese word segmentation based on language situation in processing ambiguous words. Information Sciences, 2004, 162(3-4): 275-285

[51] B. WANG, Y. Q. LI. Research on the Design of the Ontology-based Automatic Question Answering System. 2008 International Conference on Information Technology in Education, 2008,

12(5): 871-874

[52] 李慧. 词语相似度算法研究综述. 现代情报, 2015, 35(4): 172-177

致 谢

通过一年多时间的努力，毕业设计论文终于完成了。在撰写硕士毕业论文的过程中，我的导师张冬雯教授给了我精心的指导和悉心的帮助。不论是毕业设计初期的资料收集，系统设计期间的开发，以及后续的论文撰写，张老师都倾注了大量的精力和时间。张老师在指导毕业设计过程中的那种细心和负责的精神，给了我极大的鼓励，这种精神也值得我在今后的工作中去学习。在此，我要衷心地对张老师说一声，谢谢。

随着毕业设计的完成，三年的研究生学习即将结束。在这里，我还要感谢河北科技大学信息科学与工程学院的各位老师。各位老师的辛勤哺育，让我在计算机科学技术各个领域都有了很大的收获。

同时，感谢一起攻读工程硕士的各位同学，在我编写系统的过程中给我提供了许多宝贵的资料和建议。同时在我写毕业设计论文方面给我传授了很多经验和启发。

还要谢谢我的家人，家庭的爱是我努力的源泉，也是我困惑，迷茫时候的不竭动力。

最后，谨向论文评阅人和各位评委专家表示最衷心的感谢！

个人简历

李印鹏，男，1982 年 8 月出生，河北省张家口市人。2005 年 6 月毕业于石家庄经济学院电子信息工程专业。2005 年 7 月到 2014 年 3 月期间，先后在河北华恒信通信技术有限公司、石家庄众城通信技术有限公司、华为技术有限公司工作。2014 年 4 月进入石家庄工程职业学院工作至今，从事计算机通信教学工作 4 年。

曾发表过的论文有：《浅谈物联网通信技术的发展现状及趋势》、《浅谈面向 5G 的无线通信系统关键技术》等，参编了学院内部教材《计算机应用基础教程》的编写工作。