分 类 号：_____ 密 级：_____

学校代码：10414 学 号：2015010710

江西师范大学

硕士研究生学位论文

智能答疑系统的研究与实现

Research and Implementation of Intelligent Question Answering System

薛良波

院　　所：计算机信息工程学院　　导师姓名：钟林辉

学科专业：软件工程　　研究方向：软件自动化

软件演化

二〇一八年 五月

独 创 性 声 明

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

学位论文作者签名：　　签字日期：　　年　　月　　日

学位论文版权使用授权书

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

（保密的学位论文在解密后适用本授权书）

学位论文作者签名：　　导师签名：

签字日期：　　年　　月　　日　　签字日期：　　年　　月

摘　要

随着网络教育模式的出现，越来越多的教学辅导机构开始选择进行网上授课，越来越多的学生也参与到网络学习的大军中。但是，网络教学过程中，由于师生之间时间和空间的分离而不能及时的讨论交流，导致学生在学习过程中的疑难问题不能得到解答，这将很大程度的阻碍学生的学习进程，降低学生的学习效果。因此，答疑系统的开发对网络教育有着重要的作用，而建立一个能够高效快速地回答学生问题的智能答疑系统来辅助学生学习则更为重要。

本文主要目标是实现基于 PAR 方法教学网站的智能答疑系统的设计和研发，完成对程序设计这一领域相关问题的回答。首先分析了网络教育下产生的答疑系统的特点并介绍了国内外答疑系统的研究现状。然后，概述本答疑系统中所使用到的关键技术，提出隐马尔科夫提问预测模型，并通过实验进行参数设置。继而进行答疑系统的总体设计和各功能模块的设计。最后给出本系统相关关键技术的实现，并对答疑系统各功能模块进行了展示，通过对答疑系统的应用测试，验证了系统答疑的准确性。

本文的主要创新工作如下：

（1）本文研发了基于自然语言处理技术的智能答疑系统。该系统中很好地完成了对中文自然语言提问的处理，使用正向最大匹配算法进行中文分词，采用基于关键词的语句相似度计算方法和基于同义词词林的语义相似度计算方法相结合的多信息语句相似度计算方法来计算两个句子的相似程度，既考虑了语句之间的相似程度，又考虑到句子语义相似程度，使得系统能够更加准确的了解学生的需求，从而给出正确的提问回答。自然语言处理技术的引入，大大提高了系统答疑的准确率。这是创新点之一。

（2）本文首次将隐马尔科夫模型应用于智能答疑系统。首先收集程序设计领域相关问答并对这些问题进行知识点归类，获取数据集；然后使用大量提问序列进行训练得到隐马尔科夫模型，即确定各知识点之间的关联概率；最后根据学生当前提问和已知的隐马尔可夫模型来预测内在知识点的迁移，得到学生最可能的下一次提问内容，实现对学生智能推荐问题的功能。在智能答疑系统中构建隐马尔科夫提问预测模型，大大提高了系统答疑效率。这是创新点之二。

关键词：智能答疑系统；中文分词；语句相似度；智能推荐

Abstract

With the emergence of the online education mode, many teaching institutions began to lecture online, more and more students are also participating in the online learning team. Unfortunately, in the process of online teaching, teachers and students can not discuss and communicate in time because of the separation. It causes difficult problems of students can not to be solved. It also hinders students' study to a great extent and affects students' learning. The question answering system plays an important role in network education, and it is more important to establish an intelligent answering system which can answer students' questions efficiently and quickly as the students' assistant.

The main goal of this paper is to design and develop an intelligent question answering system based on the PAR method teaching website, and to complete the answer to the related questions in the field of program design. First of all, it analyzes the characteristics of the question and answer system generated in the network education, introduces the research status of the question and answer system at home and abroad. Second, an overview of the key technologies used in this question and answer system is presented, a Markov prediction model is proposed, and parameters are set through experiments. Third， the overall design of the answering system and the design of each functional module are completed. Finally, the realization of related key technologies of the system is presented. The functional modules of the question and answer system are displayed. The accuracy of the system's question and answer is verified by the application testing of the intelligent answering system.

The main innovation work of this paper lies in:

（1）This article developed a smart question answering system based on natural language processing technology. The system has successfully completed the processing of Chinese natural language questions. It uses the forward maximum matching algorithm for Chinese word segmentation and uses a keyword-based sentence similarity calculation method and a multi-information sentence similarity calculation method based on a synonym word forest-based semantic similarity calculation method to calculate the similarity of two sentences. This method not only considers the degree of similarity between sentences, but also considers the semantic similarity of sentences,

so that the system can more accurately understand the needs of students, thus giving correct answers to questions. The natural language processing technology is introduced into the answering system, which greatly improves the accuracy of the system answering questions. This is one of the innovation points.

（2）This paper applies the hidden Markov model to the intelligent question answering system for the first time. First of all, this paper collects relevant questions and answers in the field of program design and classifies the knowledge into these questions to obtain a data set. Then using a large number of question sequences to train and obtain a hidden Markov model, which means to determine the probability of association between the knowledge points; Finally, according to the students' current question and the known Hidden Markov Model, the migration of the internal knowledge points is predicted, the students' most likely next question is obtained, and the function of the system to the student's intelligence recommendation problem is finally achieved. Constructing a Markoff question prediction model in the intelligent question answering system greatly improves the efficiency of answering questions. This is the second part of the innovation point.

Key words: Intelligent Question Answering System; Chinese Word Segmentation; Sentence Similarity; Intelligent Recommendation

目　　录

1 绪论

1.1 研究背景

科技的飞速发展，使得越来越多的学生进行网上学习。网络学习就是指师生之间不受传统学习模式的限制进行学习，即学生不需要在固定的时间和空间中学习特定的内容。网络学习具有打破时间空间的限制、个性化学习、学习主体改变且互动式学习等特点。网上教学是利用网络提供的虚拟环境，学生学习与教师授课可同步或者异步进行。通过同步学习的方式，学生可以实时跟随老师的授课步伐，跟随老师的授课内容同步学习。

学生可以根据自己的时间或身处环境，制定自己的学习计划。学生的学习打破了传统的在特定时间和空间由特定老师讲授特定内容的限制，学生可以随时随地通过计算机网络进行学习、巩固和提高，不断进步。教师也能判断学生的学习进度，帮助其获取所需要的知识。传统教学的主体是老师，学生只能被动的接受老师讲授的知识，而网络教育[1-2]的学习主体是学生，学生可以主动的选取适合自己的课程。学生自主规划自己的学习时间及学习内容，根据知识的重要性合理的获取所需知识。网络学习中出现的留言论坛等平台能够为师生之间的交流提供环境，师生之间意见得到交换[3]。

网上学习的过程中，老师和学生空间上的分离，使得学生必须学会自主学习。这就要求学生能够独立思考，获取基本知识的同时还需要进行深入研究，因此答疑工具必不可少，它能够加强师生之间的交流，帮助学生获取正确的答案信息，消除学生的学习障碍。一方面，教师能够通过答疑来了解学生在学习过程中可能会遇到哪些困难以及哪些问题是学生难以理解和掌握的，老师根据学生的提问情况判断学生的整体听课水平，及时调整教学内容。另一方面，学生能够参考已有提问，避免以后犯同样的错误。利用计算机技术和网络技术解决学生在学习过程中产生的疑难问题[4]，既能节省教师时间、提高教学效率，又能统计大多数学生在学习过程中出现的共性错误，这对老师的教学重点的把握有着重要影响。

智能答疑系统[5]是网络教育的组成部分之一，比传统答疑方式更为优越且发展前景广阔。但目前的智能答疑系统普遍存在智能化程度不高、答疑不准确、缺少师生互动，这些系统存在很多不足之处。

（1）智能化不足

目前已经存在的系统，其查询方式普遍为关键词检索，且系统缺乏对学生提问进行评价的渠道，无法真正了解学生需求。

（2）自然语言提问机制不完善

使用关键词进行答疑的系统，学生需要对所提问题有一定的归纳能力，从问句中提取关键词或综合关键词组合进行提问，系统缺少对学生所提问题是否符合需求的判断，不能优化学生提问，可能导致查询结果不准确。

（3）所给出的答案不够准确

提问方式、浓缩的关键词、语句的匹配算法等，这些因素都有可能导致系统查询结果的不准确。

（4）答疑手段单一

传统的答疑系统大多使用留言、论坛或搜索匹配来进行疑难解答。该三种方式皆能实现答疑工作，但在答疑的过程中可能出现回答不及时、师生不能实时交流、搜索答案不准确等问题，这些答疑手段无疑是不全面的。

（5）答案结果呈现方式不够丰富

不管是论坛形式的答疑还是搜索形式的答疑，答案的组织结构皆过于简单，没有将结果转换成结构良好的知识[6]，答案中缺乏多媒体元素的参与。

为了解决以上问题的不足，本文研究并设计了智能答疑系统，以智能模块、助教答疑模块、名师答疑模块三种答疑方式有效的结合，共同完成答疑工作。

1.2 论文选题与研究意义

目前，课程教学中网络教学已成为其应用热点之一。在线学习在许多国家已成为国民提高自身素质的方式之一。在线教学中，答疑系统是非常重要的，其为师生交互提供一个有效便捷的平台[7]。在线答疑能够辅助网络教学，推进新型教学模式的发展。

当下，一种潮流的教学模式就是网络教育，但答疑模块在一些教学网站中并没有设计开发，学生在线学习时，遇到相关疑难问题不能及时有效的解决。大多智能答疑并不“智能”，对于学生提出的问题，系统并不能给出满意解答。本文研究的核心是基于 PAR 方法教学[8-9]网站的智能答疑系统，完成对《程序设计方法学》一书中相关问题的解答。

本系统基于自然语言处理技术并构建马尔科夫预测模型，既发挥学习者的主体作用，又体现系统的指导性，与以往智能答疑系统的设计有着巨大的不同，改善系统过于呆板的缺点。以智能答疑为主，助教在线答疑和名师答疑为辅，对问题进行准确把握，给出符合用户心意的答复，使得远程教育智能答疑系统更加人性化。智能答疑系统在人工智能技术的带动下，必将有广阔的应用前景。

本文的研究目的：

（1）智能答疑系统的构建，实现对 PAR 方法教学网站课程内容的辅助答疑，满足学生对程序设计方法学有关知识的解答。

（2）动态的知识库的构建，将有关领域知识信息以学生提问老师回答的形式，不断的统计与收集，管理员能及时对知识库进行扩充和修正，使其逐步完善。

（3）为教学网站提供一个师生互动的平台，实现师生之间信息交流和信息分享。

本文的研究意义：

（1）提高教学质量

通过学生提问，教师可了解当前学生面临的问题，确定学生的普遍理解能力和当下对学习内容的掌握程度，及时调整教学内容和教学方式。

（2）提高学习效率

网络学习环境下，学生最基本的需求就是当遇到疑难问题时，能够及时有效的解决。智能答疑系统能够将学生在课堂上的提问进行统计和存储，当学生下次提问时，能够快速解答。

（3）减少教师工作量

智能答疑系统是网络教育下教师的重要辅助教学工具。在教学过程中，学生会产生相同或者不同的疑问，教师往往需要对一些书本概念的解释和重复问题做出回答。答疑系统将简单问题及答案存入到知识库，以便学生可随时自主查找，答疑系统的使用筛选了大量简单、重复性问题，使得教师的工作量大大减少。

（4）改善教学环境

固定时间空间的教学模式，对老师的教学产生一定的影响。答疑系统的出现可伴随着学生学习的全过程，既可辅助网络教学使用也可以独立使用。学生在学习过程中遇到问题，可随时打开答疑系统进行问题解答，也可在课下自主学习时随时随地进行提问获取答案。

（5）促进远程教育的发展

智能答疑系统是以自然语言为基础，为师生交流提供平台，这对于远程教育在国内的进一步普及有着重大意义，具有很大的社会价值。

1.3 国内外研究现状

当前，网络教育十分火热，在线学习已成为学生自主学习的一种渠道，在线答疑是网络教学的一部分。而当前智能答疑系统并没有在教学网站中得到应有的重视，也没有得到相应的地位。答疑系统的研究成果较少，对系统的关键技术研究仍然有很多问题。

智能答疑系统是一种计算机程序。该程序能够对学生提问作出准确回答，将结果返回给用户。在这个过程中，能够统计学生提问的高频问题，也能够存储学生提问的新问题，提交给相关老师或专家进行回答，将新问题答案存储进知识库，实现对知识的记忆和学习。

60 年代，使用自然语言提问的智能答疑系统便产生了。随着人工智能、机器学习科学的发展，智能问答系统的研究也有了一定的突破。国内的答疑平台一般伴随着教学网站的开发而产生，作为教学网站的辅助模块帮助教学，对学生的提问进行答疑，达到巩固课上学习内容的目的。而国外的答疑系统，设计过程中采用先进的技术，能够很好的实现互动交流、准确的掌握学生提问的意图，智能化程度较高，一般可作为一个独立的平台运行。相比国内的简单答疑方式，国外的系统能够够实现对某一特定领域知识较为精准的智能回答[10]。

国外的典型答疑系统有：

（1）Start 问答系统。该系统是由麻省理工学院设计开发的，它是基于自然语言技术实现智能问答和检索的系统，也是面向互联网的世界上最早的智能答疑系统之一。系统的核心是知识库的构建，问答实现的基础也是对知识库的信息查找。该系统能够回答的知识范围包括地理、政治、艺术等领域。答疑内容丰富（包括视频、音频等），准确率较高，但是该系统不能对中文或其他语言形式的提问进行识别和理解，仅能实现英文问答，这也是该系统的一个巨大的缺陷。

（2）AskJeeves 系统。系统能够进行自然语言提问，是 AskJeeves 公司开发设计的。系统分析提问语句并能与提问者进一步交流，以便真实、全面的获取提问者的想法。但是该系统不能直接给用户提供简洁的答案，使得用户根据返回结果还需要进行信息识别和筛选，这也是该系统的巨大缺陷之一。

（3）AnswerBus。该智能答疑系统是基于搜索引擎技术支持的，系统返回给提问者的答案是综合各搜索引擎抓取的信息，将所有的获取信息都返回给用户。

在国内的教学网站中，经常可以看到网站的一个模块为智能答疑模块。教学网站将智能答疑作为其一个子功能。目前，大多的答疑系统仍主要通过人工答疑的形式与用户进行交互，如用户留言、email、在线论坛等。

（1）北京大学开发设计的智能答疑系统。在这个答疑系统中，论坛和在线讨论是答疑的主流方式。其中，论坛区域中有若干知识领域，学生根据自己的需求进入不同主题的讨论区进行询问和解惑。

（2）Vdass 在线教育平台是北京师范大学研制开发的，该平台中的智能答疑系统具有自适应的知识库。系统实现原理是对学生的提问，老师依据自己的见解，将这些问题划分为不同的知识区域，有条理的将这些问题进行存储。系统接收学生提问并立即定位到相关主题知识库智能检索，按照问题与答案的相关度大小，以一定的形式组织反馈给学生最优结果。

（3）AnswerWeb 智能答疑系统是上海交通大学研制开发的，该系统的主要特色是能够对用户提问进行动态存储，即知识库是不断扩大的。系统获取用户提问并在知识库中进行信息查找，将提问问题与知识库中每一个问题进行相似度匹配，根据匹配结果找到最类似问题，返回该最类似问题答案。在答疑过程中存在大量的人工参与，如知识库中不存在答案的回答，知识库数据信息的更新等问题。专业人员将系统不能回答的问题进行搜集和回答，若该问题为常见问题，则人工将问答添加到知识库中，使得知识库不断壮大。

1.4 论文的主要工作与结构

1.4.1 本文的主要工作

本文在查阅了大量的相关文献资料的基础上，做了如下工作：

（1）总结网络教育模式下，智能答疑系统应具有的特点，采用自然语言处理技术，研发了能对 PAR 方法教学内容进行疑难解答的智能答疑系统；

（2）首次将隐马尔科夫模型应用于智能答疑系统，提出了隐马尔科夫提问预测模型。通过分析学生提问的知识序列，对学生下一次提问进行预测，实现智能推荐功能，提高答疑效率。

1.4.2 本文的组织结构

本文的组织结构：

第一章 引言，主要介绍网络教学环境下针对学生提问而产生的答疑系统的背景，分析了当前智能答疑系统的国内外研究现状，并设定研究目标和研究内容。

第二章 概述本答疑系统涉及的相关技术。

第三章 统计程序设计领域相关问题并设计知识库，建立隐马尔科夫提问预测模型，实现智能回答模块中智能推荐的功能。

第四章 通过对答疑系统的需求分析，制定系统的总体设计计划，详细介绍系统各功能模块。

第五章 对系统各主要功能模块的展示，对系统进行应用测试，通过实验结果验证系统答疑的准确性。

第六章 总结与展望，对本文工作的总结及对未来工作的展望。

2 相关技术概述

一个具有智能性的答疑系统，应具备答疑的准确性及针对学生的提问分析学生意图给予信息反馈的特点。本文采用自然语言处理相关技术，实现系统对学生需求的准确理解，使用马尔科夫预测模型来理解学生的提问意图，准确掌握学生一次提问的目的。

2.1 中文分词技术

2.1.1 中文分词技术概述

（1）中文分词概念介绍

中文分词就是获取一个汉字序列经过切分处理，变成一个个单独的词的过程。因此对中文进行处理的首要问题就是考虑如何对中文语句进行自动分词。在英文中，单词之间以空格作为分界，而中文只能由标点和段落作为句意的分割，若想获得中文词串，无疑是个困难的问题。

中文分词的核心就是对句意的把握，即关键词的获取。传统的搜索引擎根据获得的关键词在海量的网页中进行关键信息的查找，并将最相关信息返回到搜索结果首列。中文分词的正确程度将直接影响查询结果。

（2）中文分词难点介绍

中文分词是自然语言处理技术的一部分，在词性标注、命名实体的识别、句法分析等有重要应用。本文主要实现对中文句子的分析，研究难点主要为：“切分歧义”和“未登录词的识别”。

1）歧义是指一句话经过不同的切分，产生不同的意思。主要分为：交集型歧义、组合型歧义和真歧义三种。

交叉歧义是指对字符串 AJB，AJ 和 JB 都是汉语词汇，不同的切分组和导致语句的意思也不相同。如:“化妆和服装”就可以拆分为“化妆 和 服装”或“化妆 和服 装”，这就是交集型歧义。

组合歧义[11]是指字符串 AB，同时 A 和 B 单独观察时也是词汇，这就需要根据句意来判断此句子的拆分。如:“这个门把手坏了”与“请把手拿开”，同样的“把手”就需要不同的切分。

真歧义是指只能够通过上下文信息来判断句子的拆分。如：“乒乓球拍卖完了”可以理解为“乒乓 球拍 卖 完 了”或者“乒乓球 拍卖 完 了”。

2）未登录词主要分为命名实体、新词、专业术语等词汇。这些词在分词词典中并没有收录，但在文章中或句子中会出现。最典型的出现形式就是人名。如：“王大大去上海了”，这句话中“王大大”就是一个人名，作为一个新词在句中出现。此外，还有机构名、产品名、地名、省略语、商标名、简称等都是很难处理的问题。

（3）中文分词研究现状

当前，对自然语言处理的研究中，处理中文的技术要远落后于西文处理，中文分词是很大的影响因素之一。中文分词技术在很多场景中得到应用，如机器翻译[12-13]、自动摘要[14-15]、自动分类[16-18]、语音合成[19-22]、自动校对[23-25]等，在这些应用中都需要解决中文分词问题。

2.1.2 常见中文分词方法

中文分词的研究方法有很多，本文主要介绍三种主流的分词算法。

（1）机械分词算法

机械分词方法也被叫做字符匹配方法，其实现原理是将已有的字符串与一个充分大的词典进行匹配，若在词典中找到相应的字串，则匹配成功。匹配的方式有多种，按扫描方式可以分为正向匹配和逆向匹配，按长度优先的方式可以分为最大匹配和最小匹配等。故常见机械分词方法有：逆向最大匹配法、正向最大匹配法、双向最大匹配法、最少切分法。

优点：性能高。由于匹配规则简单，只要在词典中加入新词就能很好地支持未登录词的识别，所以在工程应用上使用率高。

缺点：消歧能力较弱。规则过于简单，对词典里没有的新词不能识别。

（2）基于理解的分词方法

基于理解的分词方法指让计算机来识别词，通过模拟人的行为来理解句子。在分词的过程中，通过句法和语义的分析来处理歧义的现象。汉语言的特性决定了汉语言知识的复杂性，目前计算机对汉语言的处理还不够完善，计算机并不能充分理解汉语言信息并转换成机器可识别语言，而该方法的实施是建立在大量语言信息的基础上，故该方法目前使用较少。

（3）基于统计的分词方法

该方法是根据字符串在语料库中出现的频率来判断其是否构成词。该方法典型的算法有：CRF[26]、HMM[27]、最大熵模型[28-29]等。其中 CRF 效果更好，其比 HMM

有更弱的上下文无关性假设。

优点：该方法是基于统计算法进行切分，理论上只要有足够的训练语料，就可以很好地处理“未登录词”和“歧义”的问题。

缺点：由于该方法需要足够充分的语料进行训练，训练时间较长。工业应用中，经常会出现新增专业领域语料的情形，此时需要重新训练，较为麻烦。故一些对性能要求较高的工业应用不采用该方法。

本文主要用于对导师薛锦云教授主导教学的《程序设计方法学》课程学习答疑，其中有很多新词产生，如：PAR、PAR 方法、PAR 平台等，故使用正向最大匹配法。

最大匹配算法主要包括正向最大匹配算法、逆向最大匹配算法、双向匹配算法等。 其主要原理都是切分出单字串，然后和词库进行比对，如果是一个词就记录下来， 否则通过增加或者减少一个单字，继续比较，直到还剩下一个单字则终止，如果该单字串无法切分，则作为未登录处理。

算法流程如图 2-1 所示：

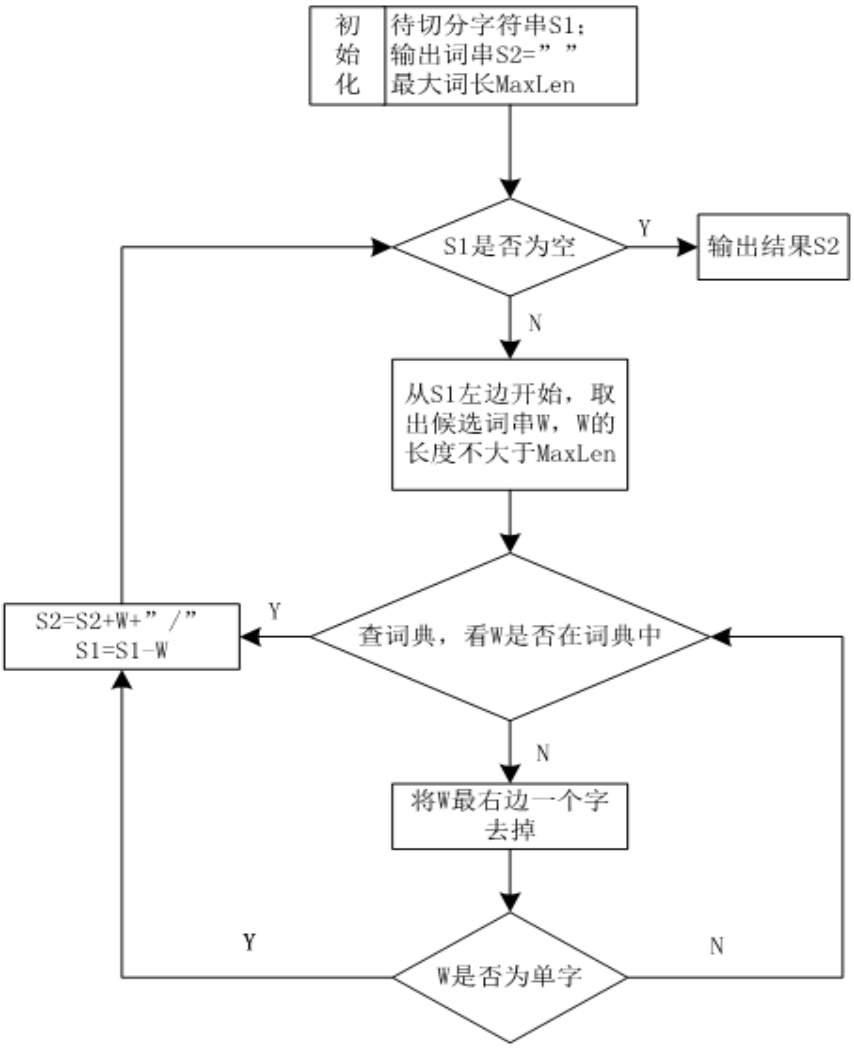

图 2-1 正向最大匹配法

2.2 语句相似度计算

2.2.1 句子相似度简介

在问答系统中时常会用到语句相似度计算[30-31]。句子相似度[32-33]是指两个句子在语句和语义上的匹配符合程度。相似度的取值为[0,1]，相似度的值越大，表示两个句子越相似，反之则两个句子语义越相离。当取值为 1 时，表示两个句子相同；若取值为 0，则表示两个不同的句子。对句子的相似度计算主要分为三类，即单纯语句的相似度、语义相似度、语法相似度。

答疑系统中的查询是指根据学生提交的问题，从问答库中检索出与学生提交问题在语义上最相似的问题，将该问题答案返回给学生。若学生能够对返回的答案给出满意的评价，则表示系统能够很好地完成答疑任务。智能答疑系统的使用节省了学生等待时间、提高了学生查找问题的效率，因此对相似问题的研究具有重要意义。当前相似问题检索面临的主要问题是解决查询问题与历史问题之间的词汇鸿沟问题[34]。相关工作中，主要研究方法为：基于话题建模的检索模型[35-36]、基于深度学习的检索模型[37-40]、基于结构建模的检索模型[41-42]、基于翻译建模的检索模型[43-44]。除了解决词汇鸿沟问题，还有相关研究[45-48]利用问题的叶子类别信息，增强相似问题检索的性能。

2.2.2 常用句子相似度计算方法

句子相似度计算是自然语言处理过程中一项基础而又核心的课程，在不同的领域或应用中都有广泛的使用。句子相似度包括语句相似度和语义相似度，常用的计算方法有如下几种：

（1）基于关键词的语句相似度计算

1）TF-IDF 方法

该方法[49-50]是基于空间向量模型 VSM，以大规模真实语料为基础，统计语料库中出现关键词词频的方法，是本文使用的计算方法。

Salton 等人在 60 年代的时候就提出空间向量模型，在 Smart 文本检索系统中首次使用到本模型。模型的主要作用是计算文本的相似度，实现原理是将中文语句向量化，以向量的空间相似度来替代相应文本语句的相似度。计算机获取两组向量，通过余弦相似度计算，返回给用户文本相似度大小。模型简单易懂且容易构建。

权重计算常用的一种方法是 TF-IDF，该方法经常被用在数据挖掘和信息检索中。这里词频就是 TF，逆向文件频率为 IDF。这是一种统计方法，方法的主要目的是判断一个个体在整体中的重要程度，如判断一个字或词在某一份文件中的重要性。大量的实验数据显示，字词出现的次数与其在文件中的重要性成正比，即字词在文件中出现的次数越多，则越重要。同时，字词出现的重要性与其在语料库中出现的次数成反比。

$$tf_{ij} = \frac{n_{ij}}{\sum_k n_{kj}} \quad \text{（公式 2-1）}$$

这里，分母表示文件中所有字词出现的个数，分子表示某一个词在文中出现的次数。如果某个词在一篇文章中频率 TF 高，在其他文章中很少出现，就认为这个词可以用来分类，具有很好地类别区分能力。

$$idf_i = \log\frac{N}{n_i} \quad \text{（公式 2-2）}$$

这里，分母表示包含检索词的文件个数，分子表示总的文件个数。则权重：

$$W_{ij} = tf_{ij} \times idf_i \quad \text{（公式 2-3）}$$

若两个句子的权重向量为：$\vec{p} = (W_{1a}, W_{2a}, W_{3a}, \dots, W_{ia})$

$$\vec{q} = (W_{1b}, W_{2b}, W_{3b}, \dots, W_{ib})$$

则两个句子的相似度用两个向量夹角的余弦值表示为：

$$\cos(\vec{p}, \vec{q}) = \frac{\vec{p}.\vec{q}}{|\vec{p}|*|\vec{q}|} \quad \text{（公式 2-4）}$$

2）词形与词序相结合的方法

该方法中语句的相似度由词形和词序相似度共同决定。其中词形相似度起主要作用。若词形相似度表示为$\mathrm{WordSim(X,Y)}$，语句长度相似度为$\mathrm{LenSim(X,Y)}$，词序相似度为$\mathrm{OrderSim(X,Y)}$，则语句相似度表示为：

$$\mathrm{Sim(X,Y)} = x_1\,\mathrm{WordSim(X,Y)} + x_2\,\mathrm{LenSim(X,Y)} + x_3\,\mathrm{OrderSim(X,Y)}$$

（公式 2-5）

其中x_1、x_2、x_3是常数，且满足x_1+x_2+x_3 =1，显然 $\mathrm{Sim(X,Y)} \in [0,1]$。该方法有着明显的缺陷。因为词形匹配没有考虑语义信息，所以可能出现语义相似的句子但句子相似度计算结果较低的现象。

3）基于编辑距离的语句相似度计算

该方法主要指计算最少的操作数目，使得原字符串能够转换到目标串。操作有

“插入“、”删除“、”替换“三种。传统的编辑距离的的操作是以字为单位，现对中文问答句子的相似度计算方法中，操作的单位为词。

（2）基于多信息的语句相似度计算。

该方法是当前答疑系统中使用的主流方法。多信息即结合不同方法的特征信息进行语句的相似度计算，因为保留了每种组合方法的优点，综合考虑语义及语句的相似性，一定程度上避免了每种方法的缺点，使得相似度计算效果更好。

句子的相似度计算不仅仅只是单纯的计算语句的相似度，同时还要考虑语义信息。语义相关度计算在很多自然语言处理的研究领域中有着重要应用，如：自动问答[51]、信息检索[52-53]、事件抽取[54-55]、词义消歧[56]、社会计算[57]等。

常见的语义相似度计算方法有以下几种：

(1)基于句法分析的句子相似度计算

通过依存句法分析可以得到句子中各成分之间的语义关系。该方法[58-62]能很好的理解句子的含义并能够得到更为准确的句子相似度。但是，当前的句法分析技术不够成熟，使得该方法的使用存在一定的误差。

(2)基于语义词典的句子相似度计算

该方法是通过定义词语间语义距离，再计算词语间语义相似度[63]来计算句子相似度。基于词林的语义相似度的计算，可借助同义词词林或知网等语义知识资源来实现。句子语义相似度计算的基础是对两个句子中所有词语之间的语义相似度计算，获得所有词语之间的语义相似度，再加权平均可得两个句子的语义相似度。

取句子 X 与 Y，X 包含的词为X_1，X_2，$\cdots X_M$，Y 包含的词为Y_1，Y_2，$\cdots Y_N$，则词$X_i(1 < \mathrm{i} < \mathrm{M})$和$Y_j(1 < \mathrm{j} < \mathrm{N})$间的相同度用$\mathrm{Similar}(X_i, Y_j)$表达。此时就得出两个句子中随意两个词汇的相同度，句子 X 和 Y 的语义相同度$\mathrm{Similar(X,Y)}$表示为：

$$\mathrm{Similar(X,Y)} = (\frac{\sum_{i=1}^{m} a_i}{m} + \frac{\sum_{i=1}^{n} b_i}{n})/2 \qquad \text{（公式 2-6）}$$

式中 ai 和 bi 的含义为：

$$a_i = \max(\mathrm{similar}(X_i, Y_1), \mathrm{similar}(X_i, Y_2), \ldots, \mathrm{similar}(X_i, Y_N)) \text{（公式 2-7）}$$

$$b_i = \max(\mathrm{similar}(Y_i, X_1), \mathrm{similar}(Y_i, X_2), \ldots, \mathrm{similar}(Y_i, X_M)) \text{（公式 2-8）}$$

本文采用带有权重信息的余弦相似度方法来计算句子间的语句相似度，采用基于词林的语义相似度方法来计算句子间的语义相似度，最后，将对句子的语句相似度（T）及语义相似度（S）做加权平均，最终得到两个句子的相似度$\mathrm{M} = \alpha\mathrm{T} + \beta\mathrm{S}$，其中 T、S 分别为语句相似度和语义相似度，α、β分别为 T、S 的权重。通过遍历问答库中的问句，得到最大的 M 值。设置一个阈值 X，若计算出来的 M 值大于阈值，则表示用户输入的问题和常用的问答库中的问题为同一问题。否则，表

示知识库中没有该问题，将该问题提交给助教老师或名师回答。

(3)基于本体的语义相似度计算

该方法[64]是利用本体来建立关键词索引，句子的语义向量通过构建句子与本体间的直接和间接语义联系来获取。

此外，还有一种基于语法的相似度[65]计算，这是一种刚刚起步研究的计算方法。日常生活中的句子结构一般不会过于复杂，不同语义的句子其语法结构往往有很大的区别，这种区别越大，在对语句相似度的识别和处理就有有利，计算结果也就会越准确。这是一个很有前景的研究方向，但目前该方向的研究较为基础，思想不够成熟，相关文献也较少，还处于起步阶段。

2.3 答案抽取

在查找到的所有可能的答案中，需要将最满足条件的答案抽取出来，需要将这些候选答案进行排序，生成最终的答案返回给用户[66-67]。答案的抽取可以分为两类：事实型答案抽取和列举型答案抽取。简单事实型问题答案的抽取一般是从相关文档中选取若干个候选答案，对这些候选答案进行排序，排序的首个答案被选为正确答案返回。列举型问题答案的抽取普遍的做法是设定一个阈值，将排序的前若干答案或分值大于阈值的若干答案作为候选答案返回。已有答案抽取方法有：基于词袋模型的简单匹配、基于表层模式匹配、基于语法结构比较[68]、基于海量数据的冗余特性、基于答案的逻辑推理验证、基于多特征的统计机器学习方法等。

2.4 本章小结

本章着重介绍了实现本答疑系统所用到的关键技术，如对用户输入的自然语言处理，其中涉及到对中文语句的分词、中文语句之间相似度的计算，答案的检索等。本系统主要将自然语言相关技术应用在自动应答系统中，提高系统应答的准确性，使得系统更加智能化。下一章将重点介绍基于隐马尔可夫的学生提问预测模型设计。

3 隐马尔科夫模型及其在推荐提问中的应用

为解决答疑平台中的相关疑难个性化推荐的问题，本章提出一种学生提问预测模型。传统的推荐算法的做法是考虑知识与知识之间的相关度，较少考虑知识随时间偏移的特征，即学生上一时刻对知识的理解将对下一时刻知识的学习产生影响。学习知识是一个不断积累的过程，也是随时间推移不断理解加深的过程。在这个过程中，一个知识的不理解将影响学生对该知识的应用及后续知识的理解。对学生提问情况进行建模，主要是对学生行为和心理的刻画。学生出错点的转移及学习知识点的变化对学生提问内容的预测有很大作用，所以对学生提问随时间偏移的动态建模十分必要。

3.1 马尔科夫模型概述

生活中，建立一个数学模型，你就可以知道对象在下一个时间点所处的状态或在一段时间后的发展趋势，该数学模型能够反映事物的变化规律，这就是马尔科夫模型。该模型的基本思想就是获取当前对象的状态信息或根据过去时间的变化规律来预测下一个时间该对象可能所处的状态集合，也能根据以往的变化规律来预测以后的发展趋势。目前，在语音识别、词性标注等领域，马尔科夫模型都有广泛的应用。

马尔科夫分析的本质是建立随机时序模型，利用统计和概率有关方法找到最有可能的路径或对象，表现为：

$$Y(i+1) = Y(i) \times P \qquad \text{（公式 3-1）}$$

其中，$Y(i)$为对象在时间 i 时所处的状态，P 指的是对象在 i 的下一时刻转换到其他状态的概率，$Y(i+1)$表示为经过一步转换后对象所处的状态。对象必须具有马尔科夫性，即对象的状态转移概率不随时间的改变而变化，这是对对象建模的首要必备条件。

马尔科夫链是一种统计模型，也是随机过程的一种。若对象的发展过程是离散且随机的，发展过程中不违背马尔科夫性，这种过程就叫做马尔科夫链。其满足以下条件：

1）对象下一时刻的状态分布只与当前状态有关；

2）状态转移概率与当前时刻的值无关。

用 $\alpha = (M, P, N)$来表示马尔科夫链模型。其中，M 表示对象所有可能的状态非空集合；状态转移概率矩阵用 P 表示；N 值表示对象初始状态概率分布。

隐马尔科夫模型也是一个统计模型。不同之处在于，模型中包含未知参数，是一个包含参数的马尔科夫过程。本文就是建立隐马尔可夫模型，通过观察值来确定参数，达到预测的效果。

用五元组$\beta = (M, Q, N, P, G)$，表示隐马尔科夫模型。其中，M 表示对象所有可能的状态非空集合；Q 表示可观察状态集；N 值表示对象初始状态概率分布；隐含状态转移概率矩阵用 P 表示；G 表示观察状态转移概率矩阵。模型中，隐含状态是不能直接观察到的，观察状态是隐含状态可能的特征体现。模型结构如图 3-1 所示：

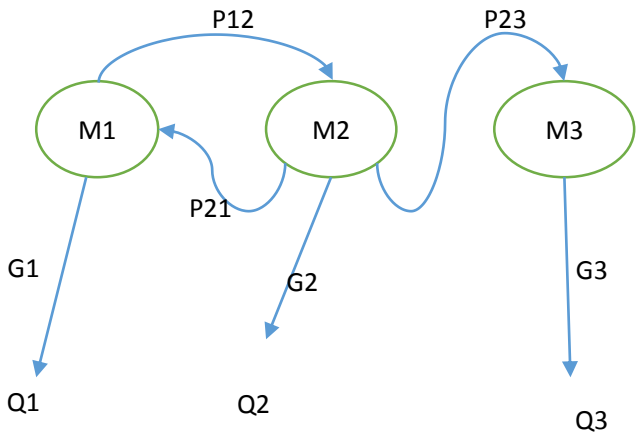

图 3-1 隐马尔可夫模型

在三种情况下可以建立隐马尔科夫模型来解决问题：

第一类问题是：给定隐马尔可夫模型，计算某一观察序列可能出现的几率有多大；

第二类问题是：给定隐马尔可夫模型，给定一组观察序列，计算产生这组观察序列最可能的状态序列；

第三类问题是：给定一组观察序列，通过调整隐马尔科夫模型来确保产生该序列的可能性最大。

当前，在学习上的预测研究主要实现对学生学习兴趣[69]、表现行为[70-73]、学习结果[74]的预测，这些预测都是对学生的最终学习结果做出的评估。而本文在学生学习过程中给学生帮助，提高学生的学习效率。

3.2 领域常用问题集的构建

在《程序设计方法学》一书中，出现很多与专业知识有关的概念、公式、定理及常见问题，本文对书中这些信息进行搜集和分类，问题的答案经过上课老师及专业教授一一查看，验证了答案的正确性，从而构建了一个有关《程序设计方法学》的答疑 FAQ 库。问答库中包括主要信息有程序设计方法学的来源、发展，结构程序设计发展及设计，程序正确性证明相关定理及证明方法，简单程序及复杂算法的程序设计方法学推导等相关知识。选取四类关系紧密且课程内容复杂的知识，对这些信息单独处理，详细规划，得到马尔科夫模型中使用的问题集。该四大知识分类分别为：程序正确性证明、简单程序的形式推导、抽象数据类型、复杂程序的证明和推导。

问题分类情况，部分列举如下：

1）程序正确性证明

a）Floyd 归纳断言

b）Hoare 公理

c）Dijkstra 最弱前置

2）简单程序的形式推导

a）赋值语句的推导

b）选择语句的推导

c）循环语句的推导

d）循环不变式的开发

3）抽象数据类型

a）数据类型的产生

b）数据类型的表示

c）ADT 的程序设计

d）基于对象的程序设计

4）复杂程序的证明和推导

a）一般程序设计方法

b）统一算法开发方法

c）算法开发实例

d）泛型程序设计

该四大问题类型分类中，每一个分类类型中有若干知识点，每一个知识点都会产生若干问题供学生提问。将程序正确性证明、简单程序的形式推导、抽象数据类型、复杂程序的证明和推导，四大类型分别编号为 A、B、C、D，具体问题的编号为 A_i_j，其表示 A 分类中第 i 个知识点的第 j 个问题，问题统计情况如表

3-1 所示：

表 3-1 问题集统计表

问题类型	知识点个数	问答总个数
A	6	25
B	5	30
C	6	43
D	7	55

将这些问题存入数据库中，得到问题集。

3.3 用户提问的动态建模

3.3.1 提问模型的确立

隐马尔可夫模型包括一般随机过程和马尔科夫链，用隐藏状态和观察序列来描述随机过程。用转移概率来描述马尔科夫链中状态的转换。用五元组$\beta=$（M,Q,N,P,G），表示隐马尔科夫模型。

（1）$M=\{M_1, M_2, M_3 \cdots M_x\}$表示在模型中共有 x 个隐藏状态，M 为状态，x 为个数；

（2）$Q=\{Q_1, Q_2, Q_3 \cdots Q_y\}$表示状态可能的观察值。y 表示观察值可能的最大数量；

（3）N 是一个概率分布，$N=\{N_i\}$表示所有初始状态的概率分布。

（4）P 是一个概率矩阵，$P=\{P_{ij}\}$，表示在 t 时刻，状态 M_i 到状态 M_j 时转移的概率；

（5）G 是一个概率矩阵，$G=\{G_{ij}\}$表示在某一时刻观察到状态 M_i 中出现观察值 Q_j 的概率。

智能答疑系统中使用隐马尔可夫模型，通过定义规则使得所建模型符合隐马尔可夫模型条件。本文提出隐马尔可夫提问预测模型，模型构建步骤如下：

（1）初步构建马尔科夫模型，明确状态、观察值代表的含义；

（2）收集历史数据，统计或定义初始状态概率分布和某状态下观察值的概率分布；

（3）根据历史数据，计算状态的转移概率；

（4）根据算法进行参数调试，使得模型趋于稳定状态；

（5）使用稳定后的模型进行预测分析。

根据提问模型的构建步骤，现在规定隐藏状态为问题分类，即本文提出的四

大分类类型；观察值表现为学生所提问的具体问题，其所属知识点可根据问题标号识别；状态转移描述为学生提问问题分类类型的迁移，如某一时刻的提问由“简单程序推导”到“复杂程序推导”的转换。本文对隐马尔可夫提问预测模型的部分概念定义如下：

定义一 提问初始状态分布概率（$N=\{N_i\}$）：表示为某一学生在第一时间选择提问的问题是 N_i 类型的概率。

计算方法为：

$$N_i = \frac{\text{提问首次提问为}M_i\text{类型的学生数}}{\text{总提问人数}}$$

定义二 提问状态转移概率（$P=\{P_{ij}\}$）：表示为学生提问类型由 M_i->M_j 转移的可能性。

计算方法为：

$$P_{ij} = \frac{\text{提问从}M_i -> M_j\text{的学生个数}}{\text{进行下一次提问的所有学生个数}}$$

定义三 观察值概率分布（$G=\{G_{ij}\}$）：表示为学生提问到类型 M_i 中知识点 Q_j 的概率。

计算方法为：

$$G_{ij} = \frac{\text{一次提问中所有询问}M_i\text{中}Q_j\text{知识点的个数}}{\text{该次所有提问个数}}$$

模型中，隐藏状态（M）和观测值(Q)可以作为已知条件，通过统计若干学生所提问题及提问顺序，得到他们的提问序列，对这些提问序列进行统计和计算，得到 N、P、G 等概率值。

3.3.2 提问模型参数的生成

本文搜集了 50 组学生提问的序列集，对该序列集进行预处理，去除无关问题和题库中未收集问题，请本书编写作者为本文中四大分类下的各知识点进行重要程度划分，该权重值即为观察值（G），概率分布情况如表 3-2 所示：

表 3-2 观测值概率矩阵

观察值概率矩阵	1	2	3	4	5	6	7
A	0.05	0.20	0.25	0.15	0.15	0.20	0
B	0.10	0.15	0.25	0.20	0.30	0	0
C	0.10	0.15	0.15	0.20	0.20	0.20	0
D	0.02	0.03	0.10	0.10	0.15	0.25	0.35

学生提问转移情况表示为：$W_x\_i\_j$ ---> $W_y\_m\_n$，如提问序列 A_1_8 → C_2_4→D_3_6→B_1_1，表示为该学生随着时间的推移，有序的提问了 A 类型的第 1 个知识点的第 8 个问题，C 类型的第 1 个知识点的第 8 个问题，D 类型的第 3 个知识点的第 6 个问题和，B 类型的第 1 个知识点的第 1 个问题。统计该 50 组提问中，初始状态下第一个问题所属的问题分类。经计算，共有 26 个学生的第一个问题为 A 类，有 15 个同学的第一个问题为 B 类，有 7 个同学的第一个问题为 C 类，有两位同学第一个问题为 D 类。故初始状态概率分布为：$N=\{0.52, 0.30, 0.14, 0.04\}$。

由定义二可得状态转移概率如表 3-3 所示：

表 3-3 状态转移概率矩阵

状态转移矩阵	A	B	C	D
A	0.10	0.25	0.05	0.60
B	0.15	0.20	0.03	0.62
C	0.01	0.04	0.20	0.75
D	0.05	0.25	0.10	0.60

3.3.3 学生提问预测

根据以上参数，获取稳定后的马尔科夫提问预测模型，如图 3-2 所示：

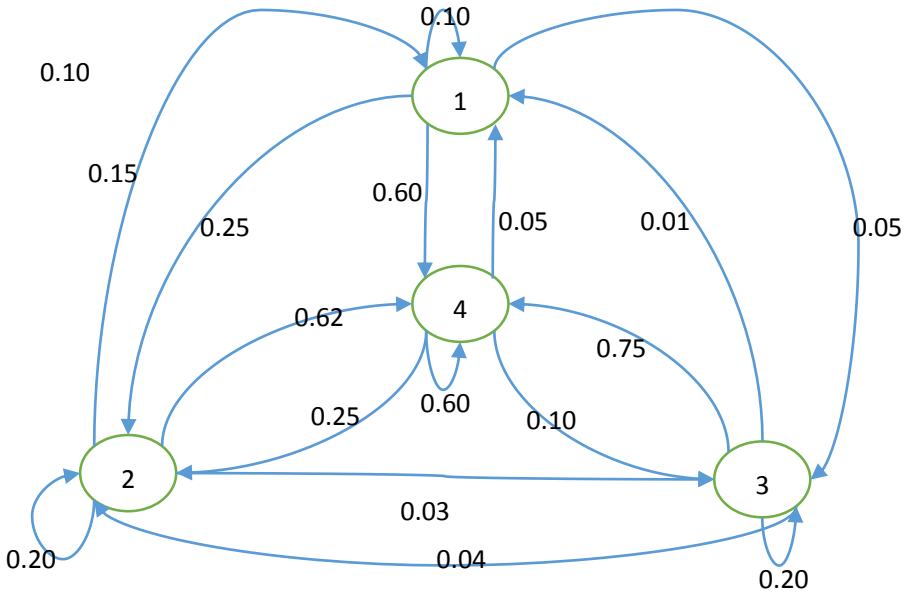

图 3-2 提问模型状态转移

由计算得到的状态转移概率和观测概率，对学生下次提问的预测过程如下：

```
begin
1 t=1
2 当前状态 Mi
3 for(j=1;j<5;j++)
4     状态转移：Mi → Mj
5         for(k=1;k<(Mj 中观察值的个数） ;k++)
6             Z[j][k]=P(Mi→Mj)*P(Qk)
7 maxP=max(Z[j][k])
8 return maxP
end
```

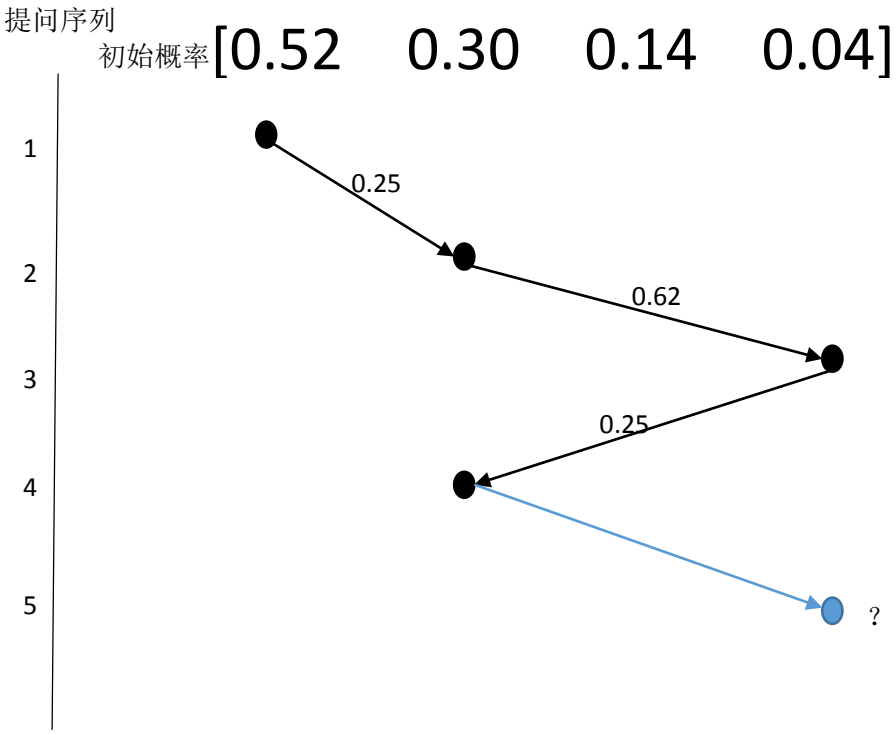

图 3-3 提问模型状态转移

根据以上提供的预测过程，很容易计算出第四次提问后，若进行下一次提问，最有可能提问 D 分类下的第 7 个知识点，其推荐概率为：

maxP=P(当前状态)*P(下一状态)*P(观察值)=1*0.62*0.35=22%

故，可以将 D 分类下的第 7 个知识点的相关问题作为推荐问题反馈给学生。

稳定后的隐马尔科夫模型，不仅可用来实现对学生提问的预测，根据学生当前提问情况，判断该学生下一个提问的可能知识点，将该知识点下所有问题选择前 5 个进行返回给该学生，作为预测提问推荐给学生。此外，根据用户的提问序列，可以很便捷找出隐藏状态之间的转换，即知识点的关联程度及知识点的衔接情况。维特比算法就是用动态规划解释隐马尔科夫模型预测问题，即用动态规划求概率的最大路径。

维特比算法：

输入：模型$\gamma = (\mathrm{P}, \mathrm{G}, \mathrm{N})$和观测 Q =（$q_1, q_2$，…，$q_T$）；

输出：最优路径$I^* = (i_1^*, i_2^*, \ldots, i_T^*)$.

(1)初始化

$$\delta_1(\mathrm{i}) = \pi_i b_i(o_1), i = 1,2,\ldots,N$$

$$\psi_1(\mathrm{i}) = 0, i = 1,2,\ldots,N$$

（2）递推. 对 t=2, 3，…，T

$$\delta_t（\mathrm{i}） = \max_{1 \le j \le N}[\delta_{t-1}(j)a_{ji}]b_i(O_t), i = 1,2,\ldots,N$$

$$\psi_t(\mathrm{i}) = \arg\max_{1 \le j \le N}[\delta_{t-1}(\mathrm{j})a_{ji}], \ i = 1,2,\ldots,N$$

（3）终止

$$P^* = \max_{1 \le i \le N}\delta_T(i)$$

$$i_T^* = arg\max_{1 \le i \le N}[\delta_T(\mathrm{i})]$$

（4）最优路径回溯. 对 t=T-1，T-2, …, 1

$$i_t^* = \psi_{t+1}（i_{t+1}^*）$$

求得最优路径$I^* = (i_1^*, i_2^*, \ldots, i_T^*)$。

如：某学生提问如下：1_7—>3_6—>4_2。表示学生提问了第一个知识点的第 7 个问题，第三个知识点的第 6 个问题，第四个知识点的第 2 个问题。现求该学生的提问最可能涉及到怎样的知识类型转换。

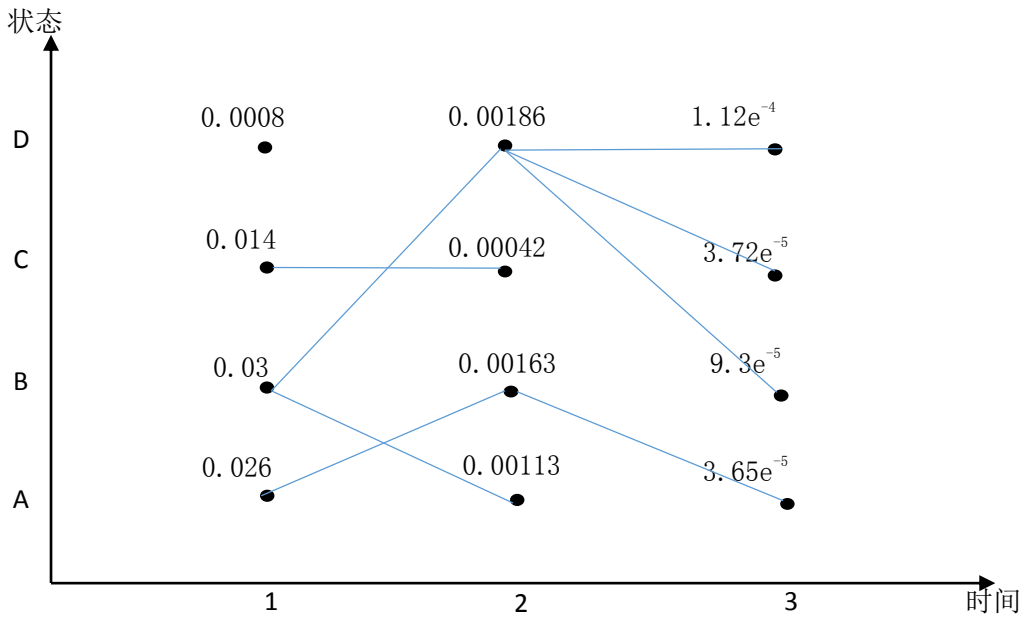

图 3-4 求最优路径

由图 3-4 可以看出，对该学生的连续三次提问，

t=3 时，最优的终点$i_3^* = argmax[\delta_3(i)] = D$

由最优路径终点i_3^*，逆向找到i_2^*，i_1^*；

在 t=2 时，$i_2^* = \psi_3(i_3^*) = \mathrm{D}$

在 t=1 时，$i_1^* = \psi_2(i_2^*) = \mathrm{B}$

于是求得最优路径，最优状态序列$I^* = (i_1^*, i_2^*, i_3^*)$=(B, D, D)。即该学生最可能依次分别提问了 B 类知识、D 类知识、D 类知识。

根据学生出错序列，找出知识转换过程中学生难以理解的地方，这对教师讲课内容重点的划分有着指引作用。同时，根据提问隐藏状态的转换概率的大小，侧面验证哪些知识的连接是更为密切的，这对教师教学内容的制定有着重大参考价值。

3.4 本章小结

本章针对智能答疑系统中智能推荐功能，统计并分类程序设计相关问题，罗列相关知识点下可能的所有问题，构建问题集。提出了马尔科夫提问预测模型，收集学生提问知识序列，设计实验进行参数调试，给出学生下一次提问的问题预测方法。下一章将重点介绍对系统的详细设计。

4 系统分析与设计

教学或自学过程中不可避免的出现学生会遇到疑难问题的情形，若由老师来回答学生遇到的所有问题，将占用老师大量的时间和精力，从而在很大程度上影响老师的教学质量和教学效率。若提取出一些非必须老师回答的问题（如：学生重复提问问题、书本中常见概念性问题、常见例题的解答等），这将大大提高老师的教学效率。基于辅助课程教学的目的，对本答疑系统进行了详细设计。

4.1 系统的需求分析

4.1.1 功能需求

本智能答疑系统的主要设计目标是对学生提问进行回答及对知识库的丰富。主要答疑方式为智能模块的答疑、助教在线答疑和名师答疑。主要用户角色分为学生用户、助教老师用户、名师用户和后台管理员用户，分别实现学生提问、概念及定理信息的智能回答、助教在线常见课程问题回答、名师疑难回答和后台管理员知识库的维护及管理。答疑系统可分为前台信息展示和提交以及后台信息管理和维护两个部分。在详细设计过程中，应考虑如下问题：

（1）查询功能。答疑的本质就是信息检索，答疑系统能快捷准确的找到提问的答案。

（2）问题进展跟踪。一个问题被提出后，必须应该得到圆满的解决，而不是长时间得不到回答或者问题提出后便“杳无音讯”。

（3）答疑评价。对于系统的回答，学生根据自己的理解或判断，对回答进行评价，若回答不满意，可提出请求以引起老师对该问题的重视。

（4）回答的更正。对于学生不满意的回答，老师需检查或查阅资料判断回答的准确性。若回答不准确，则立即更正。

（5）知识库的更新。不同的学生在学习中遇到的问题也不相同，不断的更新和完善知识库信息，这对问答的质量提升有着重大意义。

智能答疑系统使用的基本流程为：学生用户登录后，可分别使用三个模块进行提问，也可以由智能答疑模块开始提交问题并对回答进行评价，若对答案不满意，可依次对助教老师和名师提问。助教老师登录后，查看学生在线提交的问题，

以实时对话的形式对提问进行解答。名师登录后，查看学生留言问题并对这些问题一一解答。三个模块的协作回答使得本答疑系统取得更好的答疑效果。

通过对系统进行需求分析，从用户登录角度可以将本系统分为四大模块：学生模块、助教老师模块、名师模块和管理员模块。分别完成提问、常见问题回答、疑难问题回答、用户及知识库信息管理等功能。不同的用户登录系统即可完成不同的模块。

（1）学生用户

学生用户使用智能答疑系统主要目的是提问并及时获取答案，本智能答疑系统允许用户使用自然语言进行提问，当学生提出问题时，系统能自动到知识库中寻找答案，并将相似度大的若干条回答根据相似程度由大到小依次排列，返回最相似问题答案，学生根据需求判断系统给出的答案，并最终给出对答案是否满意的评价，如果不满意则系统将问题提交给助教答疑模块，由助教老师进一步解答。

学生用户用例图如图 4-1 所示：

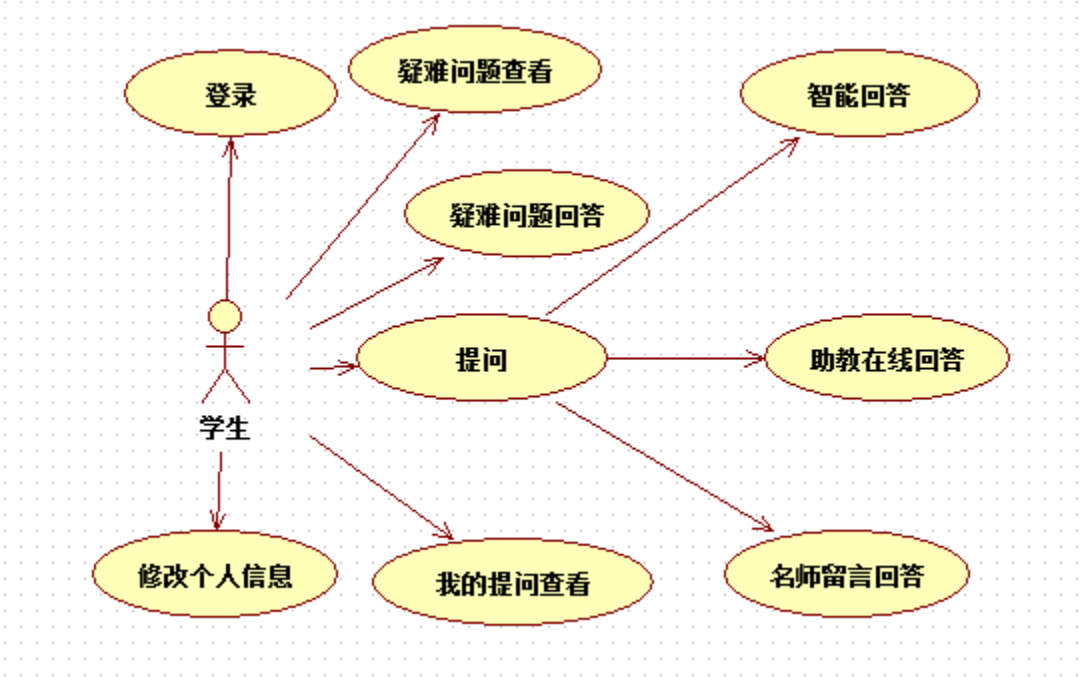

图 4-1 学生用例图

（2）助教老师用户

助教老师一般具有丰富的专业知识，主要回答系统中暂时没有存入的问题和解答。教师回答完该问题后，系统自动将问题发送给学生并同时添加到知识库中。在助教老师成功登录并进入系统后，系统会提示其查看在线用户的提问。此外，在用户提问后，系统将记录所有学生用户的提问，助教老师查看这些问题并将其中较为反复和常见问题进行统计回答，然后将这些问题及答案存入知识库中，实现知识库的不断更新和丰富。

助教老师用户用例图如图 4-2 所示：

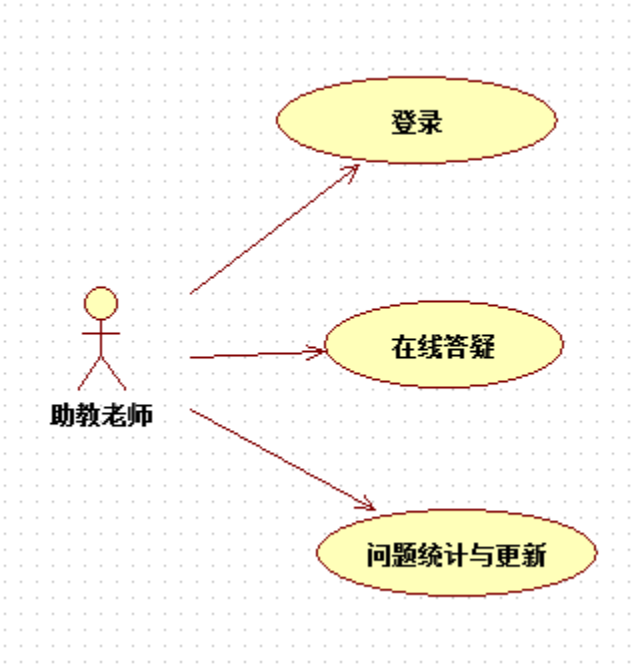

图 4-2 助教老师用例图

（3）名师用户

名师一般指在专业领域具有独特贡献及在专业知识上有独特见解或深入理解的专家学者。名师的指导对学生往往有醍醐灌顶的作用，同时，名师也能时常发布一些疑难问题供学生思考，开发学生的思维。

名师用户用例图如图 4-3 所示：

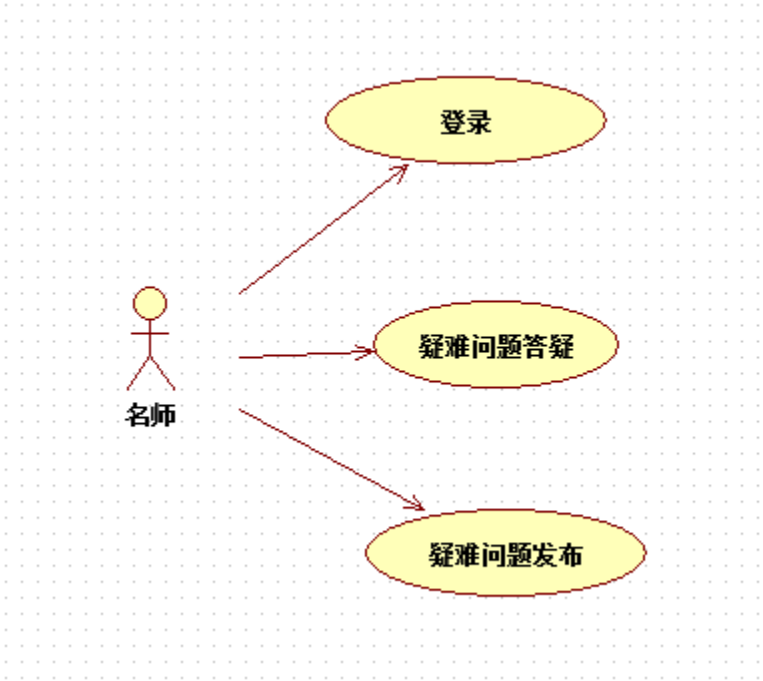

图 4-3 名师用例图

（4）管理员用户

管理员用户管理系统的原始数据，具有最高的管理权限，一般不参与对学生提问的回答，主要对系统中知识库和登录系统的用户进行管理和维护。

管理员用户用例图如图 4-4 所示：

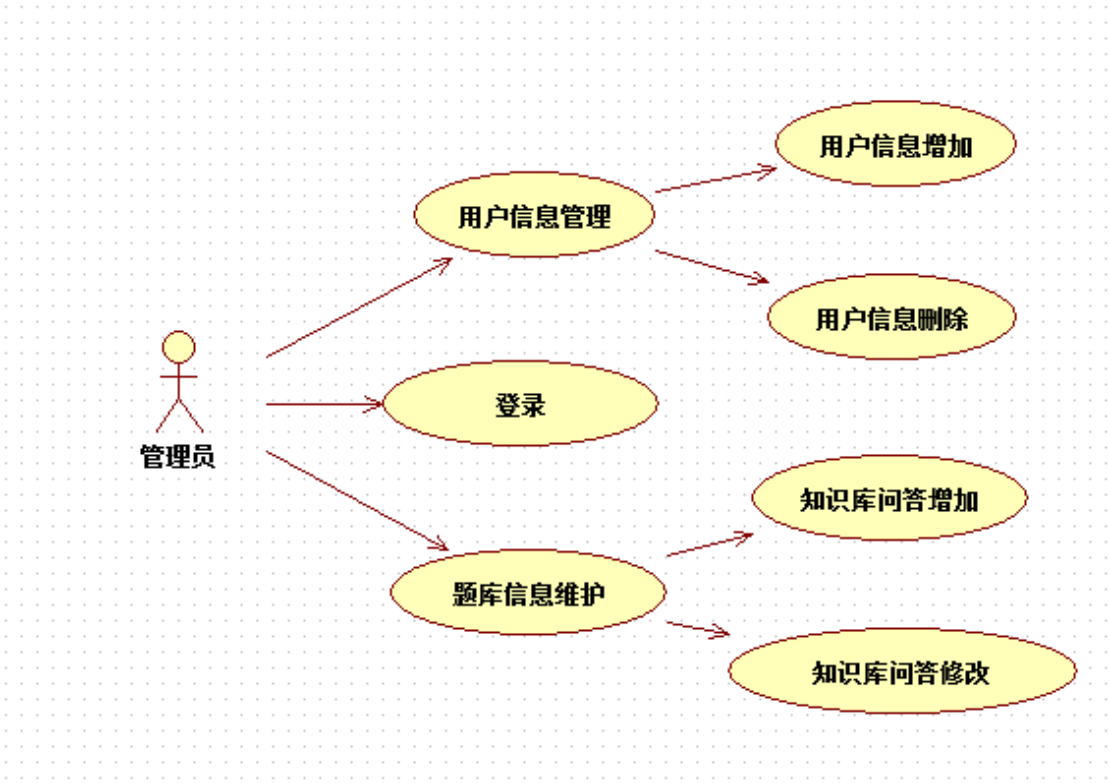

图 4-4 管理员用例图

4.1.2 可行性分析

可行性分析的目的是用最小的代价在尽可能短的时间里确定问题是否能够解决。可行性分析的本质是对系统分析和设计的简化过程，即用抽象的方式进行分析和设计的过程。本智能答疑系统虽然用户量庞大，但是设计目标明确。本文从以下三个方面进行可行性研究：

（1）技术可行性分析

使用当前主流的应用程序开发工具及技术。本智能答疑系统是基于 JavaWeb 工程开发的项目，采用 Mysql 数据库存储问答信息，适应当前主流的开发语言及技术。这些技术都是成熟可靠的，所以从技术层面考虑，本系统的开发是可行的。

（2）经济可行性

本系统开发工具为 MyEclipse，开发过程中只需要一台 PC 机。系统开发的成本为项目开发费用及运行维护费用，经估计，开发成本不高。所以在经济上也是

可行的。

（3）操作可行性

本系统的操作对象为学生用户、助教老师、名师及系统管理员。本系统采用图形界面的形式展示，界面美观大方形象，操作十分方便。非常适合普通人群使用，对于错误操作本系统会有相应的错误提示，简单、易学、易用，故本系统操作可行。

4.2 系统总体设计

4.2.1 系统架构

智能答疑系统作为网络教学系统的一部分，在设计上要遵循以学习者为中心的设计理念，智能答疑系统应该能实时、准确的解决学生在学习中遇到的问题。智能答疑系统的设计需要注意以下几个准则：

（1）以学生为中心，系统在设计过程中应注意把握学生的特征，能够理解学生的自然语言问题，判断学生的意图，给予学生正确的学习引导，达到启发性答疑的效果。

（2）激发学生的学习动机。系统应能够不断地吸引和引导学习者想学习爱学习的热情，让学习者参与到整个学习和答疑的过程中，发挥他们的主动性。

（3）答案反馈的及时性。当系统选择智能答疑模块进行提问时，系统应快捷且准确的给予学生答案，避免学生的长时间等待。

（4）知识库的更新和完善。知识库是答疑系统的核心，直接影响着答疑系统的性能。一个完善的答疑系统应该是能够不断的更新和信息补充的。

一个功能完整的智能答疑系统，应该保证尽可能的完成对各种学生提出的问题的解答，同时提供多种答疑方式。随着知识库的不断扩充，自动答疑的命中率越来越高，使得答疑活动越来越高效。基于以上的设计原则，本智能答疑系统结构如图 4-5 所示：

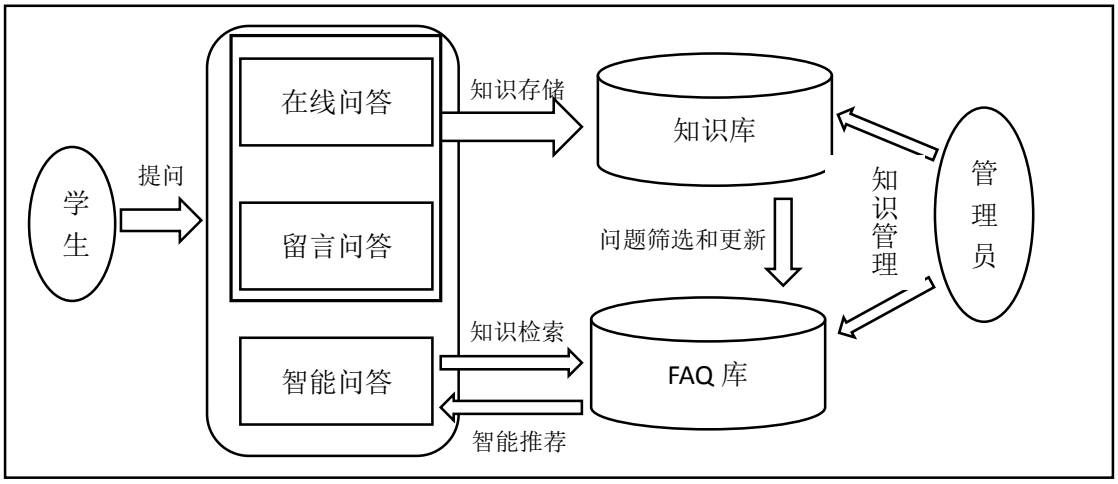

图 4-5 系统构架图

4.2.2 系统的功能结构

本文对智能答疑系统的研究主要致力于提高对问题回答的准确性、及时性、智能性以及系统各功能模块之间的工作协调性。对学生提出的问题，系统能够及时给与准确答案或者将问题提交给助教老师或者名师进行回答。智能模块的及时准确回答、助教老师的详细追问获取学生的准确需求以及名师对问题的精致讲解，各答疑模块可相互独立又相辅相成共同构成了本智能答疑系统。

智能答疑系统的基本流程如下：

（1）学生登录进行提问，系统获取提问并与知识库进行相似或相近问题的答案匹配，若搜索到相似问题及答案，则返回该信息；若无相近信息，则返回“题库无此问题答案，请寻求助教老师帮助”。

（2）学生可直接登录进入助教在线答疑模块或在智能答疑模块中问题没有得到满意解答时进入该模块，输入自己提问与在线老师进行互动交流。

（3）学生可直接登录进入名师答疑模块或在助教答疑模块中问题没有得到满意解答时进入该模块，在留言板中对名师进行留言。

（4）在学生个人中心，可点击“我的提问”，查看该用户的所有提问及回答，对系统给出的问题回答做出评价，以方便管理员对问题库的更新和修改。

参与使用智能答疑系统一共有四种角色，分别为学生用户，助教老师用户、名师用户、系统管理员用户。四种用户在系统的使用过程中，分别有自身的特点和需求。本系统分为前台操作和后台管理两个网站，前台网站主要由学生、助教老师、名师登录分别完成各自的需求及任务。后台网站则由管理员登录完成对用户和知识库信息的管理。智能答疑系统功能结构如图 4-6 所示：

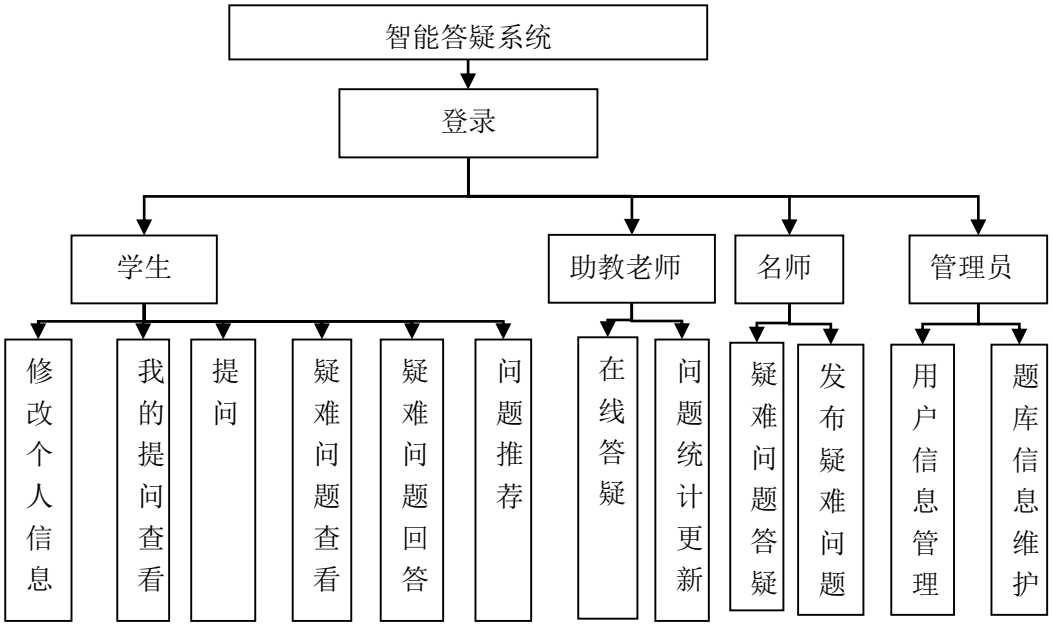

图 4-6 系统功能结构图

4.3 数据库设计

数据库是数据库应用程序的核心。数据库的设计是建立一个应用程序最重要的步骤之一。经过对系统的需求分析及功能设计，系统中的实体类型有：学生、助教老师、名师、管理员、问题、答案等。系统采用 MySql5.5 作为数据库管理系统，数据库名称为：db_autoAnswer。各表结构如下所示：

表 4-1 知识库表（questions）

field	type	default	pk	not null	auto incr	comment
id	varchar(20)		√	√	√	问题编号
qname	varchar(100)			√		问题
answer	varchar(100)					答案
number	int(11)			√		访问次数
level	int(5)			√		级别
category	varchar(20)			√		类别
multimedia	Blob					多媒体材料

questions 表用于存储所有具有标准回答的问答对，即存储的信息为知识库内容。除了基本的问答之外，该表中还存储了对该问题的分类，问题的回答可以是多种多样的，故对问题的回答还存在文本文件形式和视频文件形式。

表 4-2 用户表（user）

field	type	default	pk	not null	auto incr	comment
id	int		√	√	√	用户 id
username	varchar(20)			√		用户名
password	varchar(20)			√		密码
sex	varchar(5)	男				性别
email	varchar(20)					邮箱
telephone	varchar(20)					电话
introduce	varchar(50)					个人介绍
role	varchar(5)	学生				角色
registTime	Date					注册时间

user 表用来存储所有注册的用户信息，用 role 来区分注册用户的类别为老师或是学生，默认为学生用户。

表 4-3 用户问题表（user_question）

field	type	default	pk	not null	auto incr	comment
id	int		√	√	√	提问编号
qname	varchar(100)					问题
answer	varchar(100)					答案
states	varchar(20)	0				问题状态
satisfied	varchar(20)	满意				是否满意
time	Date					回答时间
teacherID	varchar(20)					回答老师
userID	int					提问者

user_question 表记录所有学生用户的提问并对问题进行跟踪。问题状态为“已回答”或“待回答”，学生对老师回答的评价分为“满意”及“不满意”，对于不满意的回答，可由名师进行最终解答。

表 4-4 疑难问题表（problems）

field	Type	default	pk	not null	auto incr	comment
id	Int		√	√	√	难题 id
title	varchar(20)					难题类型
professorID	Int					名师 id
ftime	Date					出题时间
context	varchar(20)					难题内容

problems 表是为名师用户设计，名师用户登录后，可进行疑难问题的设置，加强师生之间的互动，同时也促进学生学习，提高学生学习兴趣。

表 4-5 疑难问题回答表（p_answer）

field	type	default	pk	not null	auto incr	comment
titleID	int		√	√		疑难 id
userID	int		√	√		学生 id
p_answer	varchar(50)					疑难回答
p_time	Date					回答时间

p_answer 表为学生对名师设置的疑难问题进行回答。通过对疑难问题的思考，可以加深学生对某一知识点的深度研究。

4.4 系统模块设计

4.4.1 登录模块

本系统的用户角色共有四类，由于拥有的权限和实现的功能不同，用户登录将进入不同的主页面。通过用户输入的信息（包括用户名、密码、角色）进行登录验证，对合法用户进行权限分配。登录模块结构如图 4-7 所示：

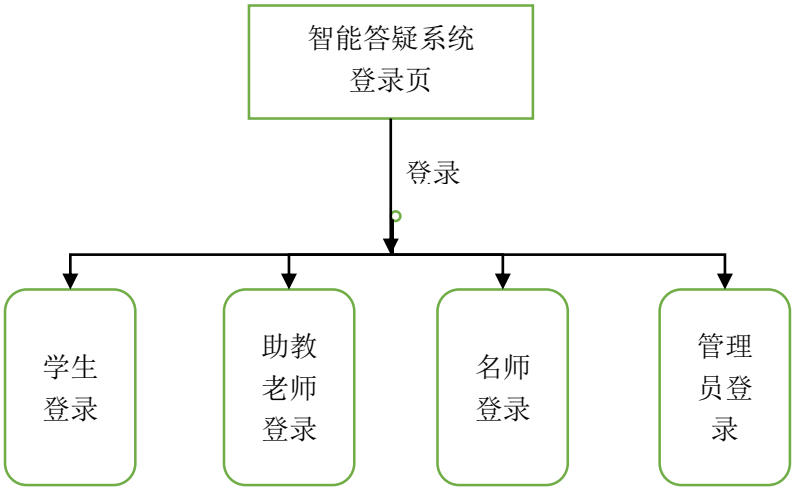

图 4-7 用户登录

4.4.2 智能回答模块

智能回答模块由四个部分组成：问题理解、智能搜索、问答抽取和智能推荐。智能回答模块结构如图 4-8 所示：

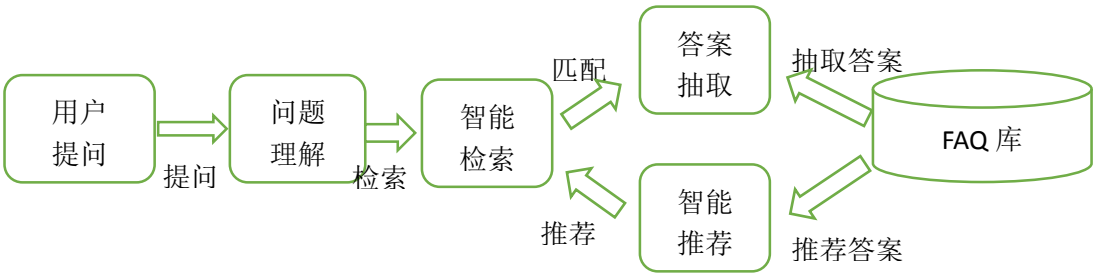

图 4-8 智能问答模块结构图

（1）用户提问

学生用户登录后，使用自然语言或关键字进行提问。

（2）问题理解

问题理解，也叫问题分析，主要处理用户使用自然语言提出的问题。它的处理过程是：先获取用户的自然语言提问，再对提问的句子进行中文分词；除去分词后词组中停用词和疑问词，保留问句中关键词；对关键词做词性标注。问题分析的主要意图是将学生用户提出的自然语言问题进行预处理，使其成为计算机能够识别并处理的格式。

（3）智能检索

智能检索的主要意图是根据之前对问题的预处理，将得到的关键词序列与知识库中问题进行匹配检索，通过语句相似度计算及语义相似度计算，获取知识库中所有相似问题，并将这些问题由相似度大小进行排序。

（4）答案抽取

问答系统的返回结果各不相同，如一大堆相关内容的网页或网页链接或搜索出若干条答案供用户参考。本文智能问答模块返回的结果是用户提问后对知识库检索排序得到的最相似问题答案，答案准确而简洁。

（5）智能推荐

一个具有智能化的答疑系统能够判断学生的意图，从而进行相关学习内容的推荐，使得系统的答疑过程更加高效。

4.4.3 助教回答模块

助教老师的主要任务是与学生在线实时互动并将学生提问进行整理添加到知识库。学生登录后对助教老师进行提问，等待老师作答；助教老师登录并查看问题，根据问题作答并将答案实时发送给学生；学生接收到答案并对老师的回答进行评价；老师根据对课程知识的理解判断该问题是否应该被收录到知识库中。助教回答模块结构如图 4-9 所示：

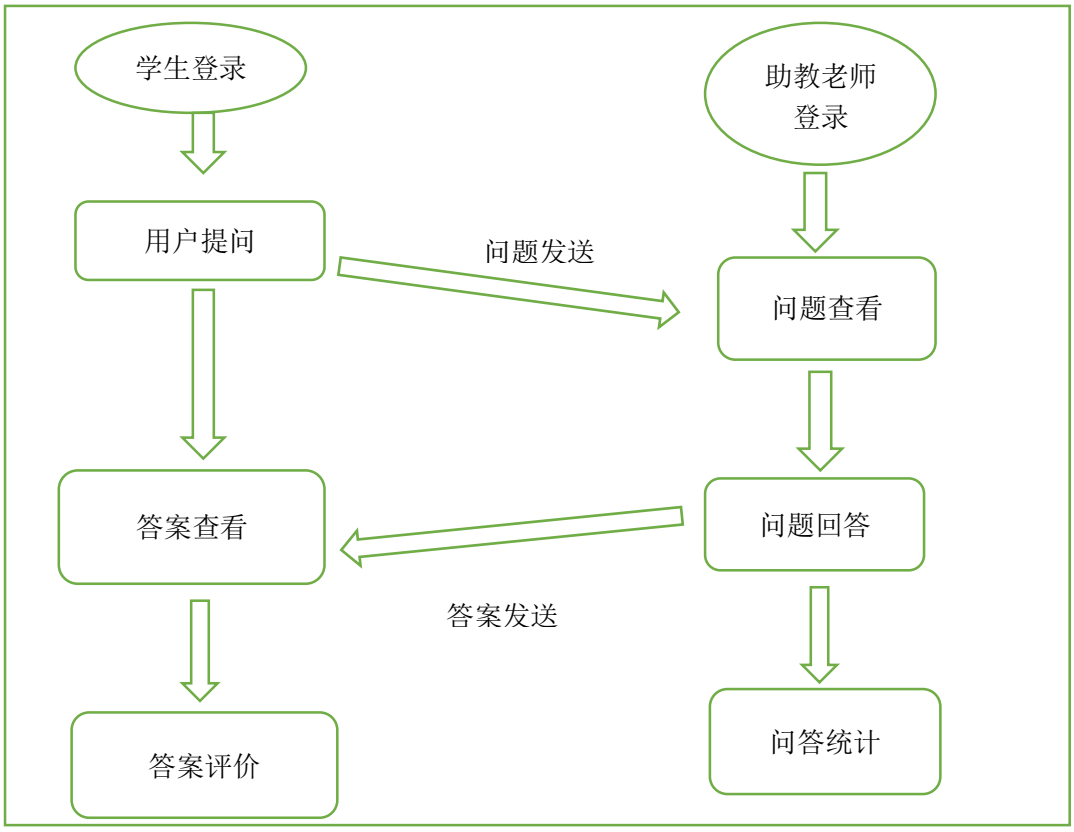

图 4-9 助教在线答疑

4.4.4 名师回答模块

名师的主要任务是进行疑难解答和疑难问题的发布。学生登录并对名师进行疑难提问留言，等待名师疑难讲解；名师登录后，查看学生留言并进行解答；学生再次登录可查看名师解答情况。名师回答模块结构如图 4-10 所示：

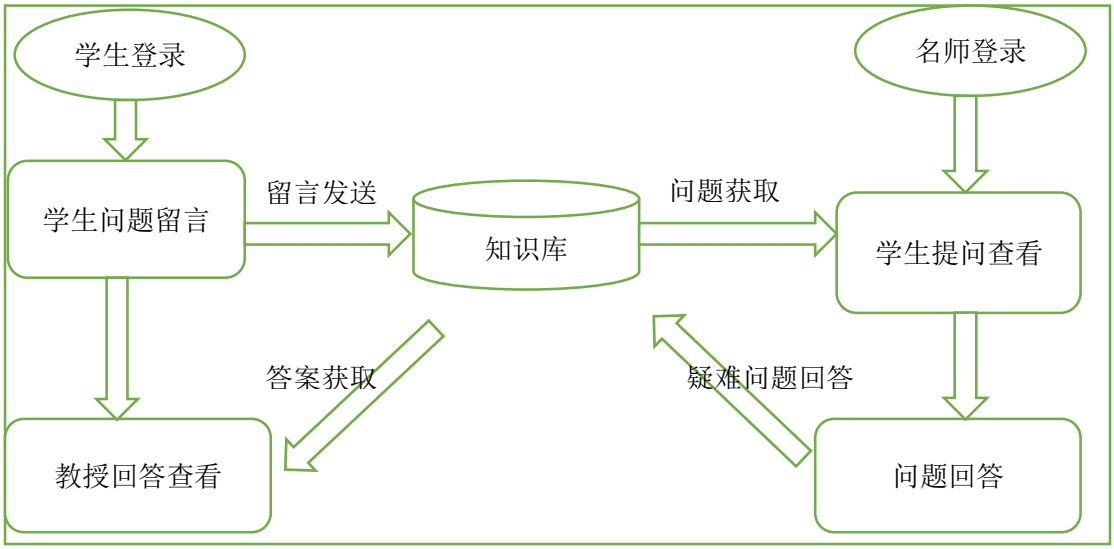

图 4-10 名师答疑

4.5 系统流程设计

本系统共有 4 类用户，不同的用户需求不同，故应该完成的功能也不相同。每种角色登录后的操作流程详细介绍如下。

（1）管理员用户流程设计

管理员用户访问本系统，输入用户名密码后成功登录到后台系统首页，可实现用户信息管理及题库信息管理。即用户注册后会将注册信息保存在该数据库中，管理员可实现对用户的添加和信息修改，当用户不再使用本系统并想要销毁个人信息时，管理员可在后台数据库中删除其信息。同时，有新的经典问题产生，管理员也可以直接将该问答添加到知识库中，无需助教老师对问答的审核。

管理员用户登录使用流程如图 2-11 所示：

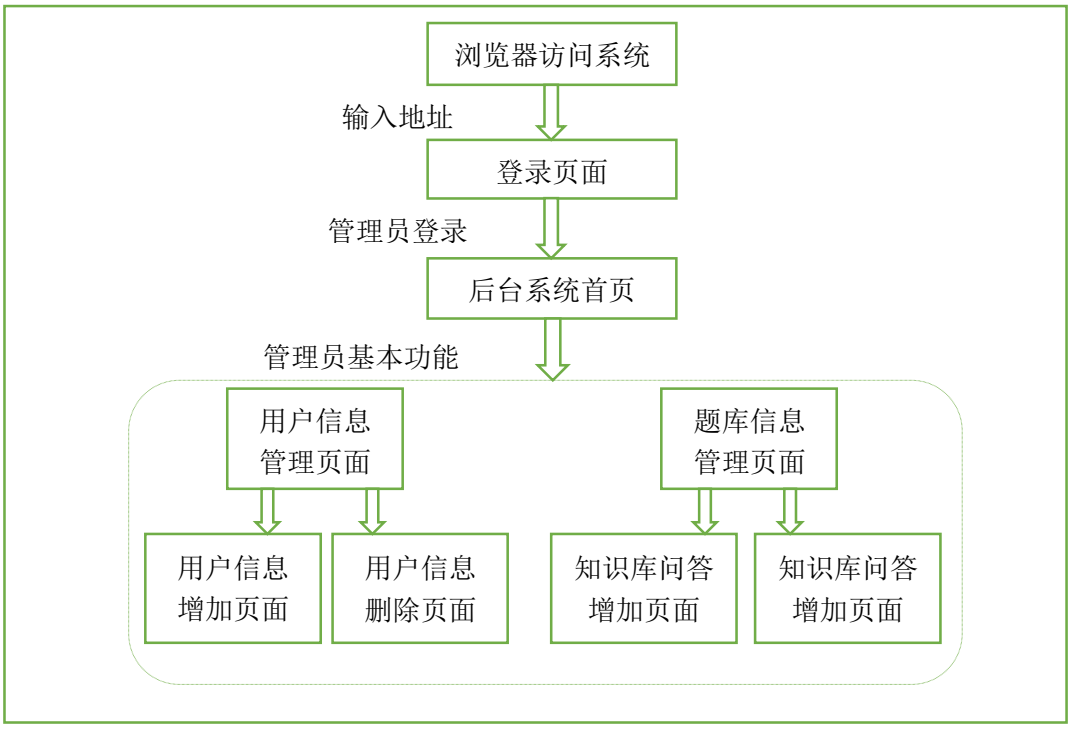

图 4-11 管理员用户流程

（2）学生用户流程设计

学生用户登录系统后，可进行四种操作：

1）进行提问。提问是学生使用本系统的是要目的。系统设置三种回答模式，即智能回答模式、助教老师在线回答模式、名师答疑模式。三种答疑方式可相互独立完成答疑工作，也可相互辅助，共同完成对学生提问的回答。

2）我的问题查看。对于学生的提问，本系统会实时跟踪问题的回答进展情况。如“问题待回答中”或者“回答不满意”或“完美解决”。

3）问题回答评价。对智能模块或助教老师答疑模块的回答，学生判断该回答的准确性，若不满意或不理解问题的回答，可选择“不满意”，系统将记录该问

答，由助教老师查阅资料判断答案的准确性或交由名师裁决。

4）名师发布的问题查看与解答。

学生用户登录使用流程如图 4-12 所示：

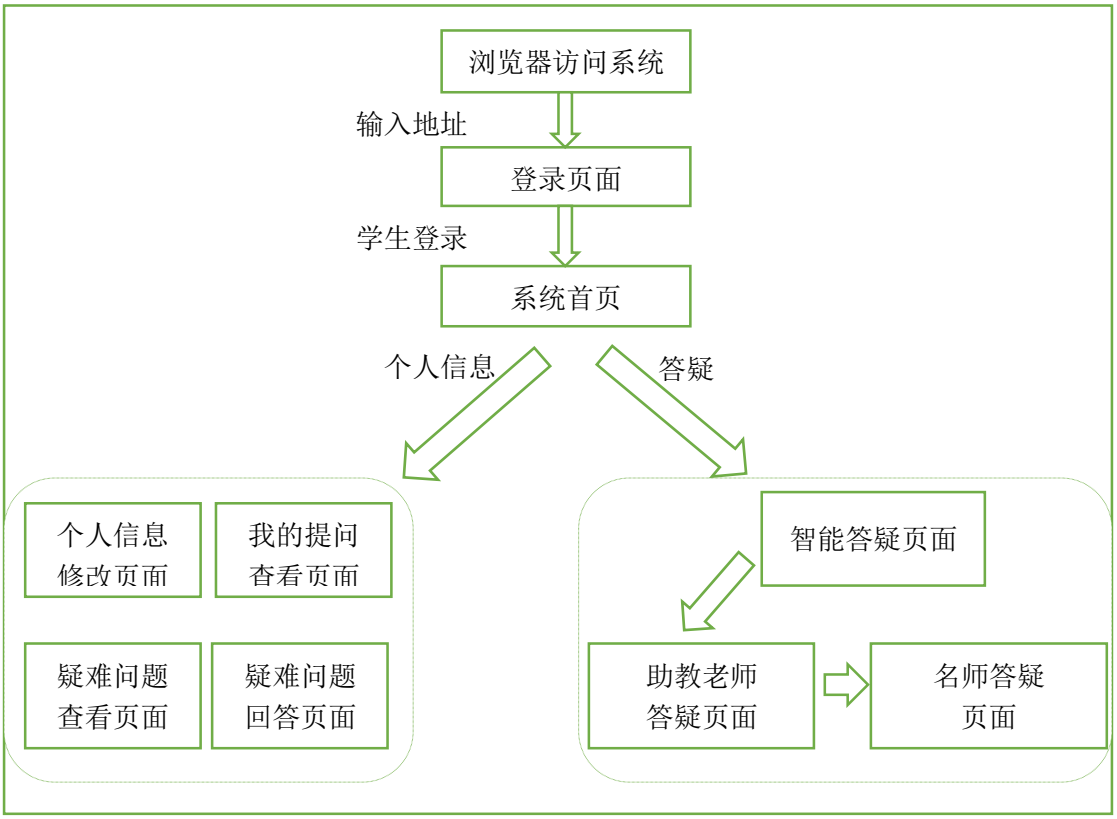

图 4-12 学生用户流程

（3）助教老师用户流程设计

助教老师的主要职责是对智能模块未成功解决的学生问题进行再次解答，对学生遇到的课程内容的问题解答，并记录这些问题，为知识库的更新提供数据基础。助教老师登录后，能立即查看学生发来的问题，老师与学生进行实时在线互动，以聊天的形式便轻松地解答困扰学生的问题，为学生提供一个友好愉快的答疑环境。

助教老师用户登录使用流程如图 4-13 所示：

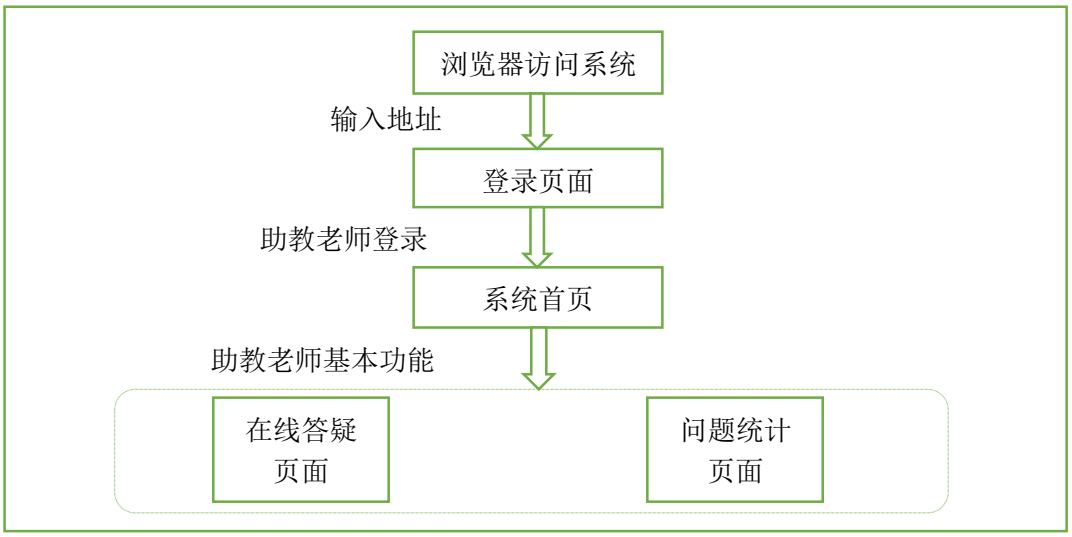

图 4-13 助教老师用户流程

（4）名师用户设计

名师，即领域知识专家，其主要职责是进行疑难问题的解答，名师业余时间可发布难题给学生思考，激励学生深入学习的同时也大大提高了学生学习的积极性。名师登录后即可查看所有学生疑难问题，名师可对这些问题进行权威解答，反馈答案给学生，供学生再次登录时查看。

名师用户登录使用流程如图 4-14 所示：

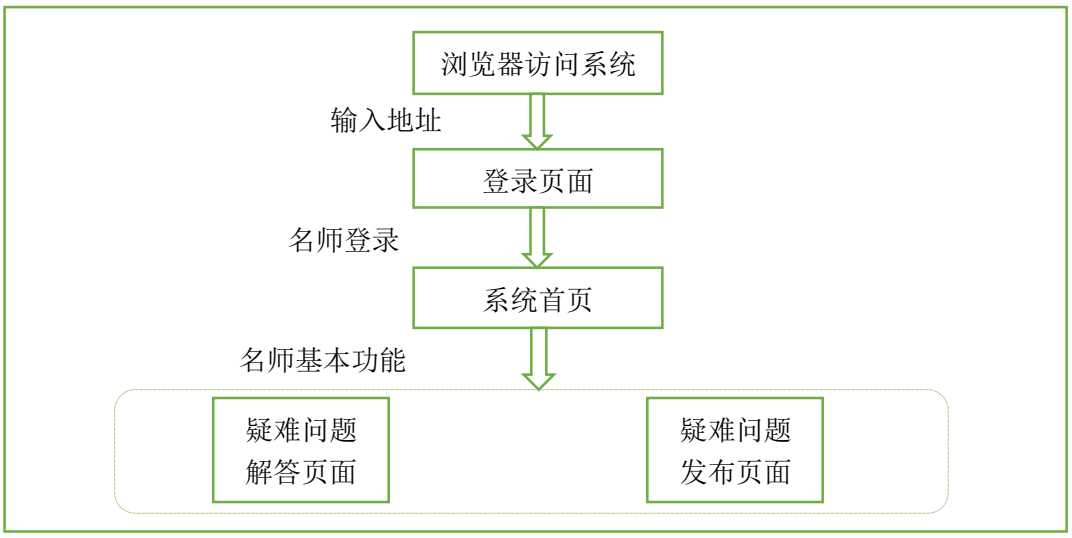

图 4-14 名师用户流程

4.6 本章小结

本章对智能答疑系统的设计进行了详细介绍，包括系统的需求分析、系统的架构设计、系统数据库设计、系统各功能模块设计及系统业务流程设计等。本智

能答疑系统答疑功能分为智能答疑、助教在线答疑、名师答疑。智能答疑用于回答学生遇到的概念、公式等问题，助教老师在线回答程序设计方法学课程内容相关问题，名师答疑主要解决程序设计领域疑难问题的回答。本系统共有四类用户，分别为学生，助教老师、名师及系统管理员。其中，学生可使用本系统进行疑难解答，助教老师和名师主要负责疑难问题回答，管理员用户负责后台知识库的维护，各角色之间相互协作，凸显本智能答疑系统的优越性。下一章将重点介绍系统关键技术的实现及对系统主要功能实现的展示。

5 系统关键技术与主要功能的实现

上一章是对智能答疑系统的总体设计，本章将具体实现系统设计提到的相关技术及功能。一个具有智能化的答疑系统应能够准确理解学生自然语言表述的问题需求，系统对学生提问作出回答的同时还应自动化给出学生下一个需求的相关推荐。本章根据系统功能实现所需要的理论技术进行具体研究和实现，主要有：分词技术、相似度匹配技术等，设计系统的详细解决方案，并对系统设计提到的重要功能进行展示。

5.1 系统主要技术的实现

5.1.1 中文分词的实现

在智能答疑模块中，对用户的自然语言提问，系统需先获取该提问并对提问进行中文分词。中文分词是指将中文语句进行句子拆分，得到一个个独立的中文词语。该过程是将中文语句根据制定好的规则拆分成中文词序列的过程。

在本系统中实现了基于词典的正向最大匹配算法。

代码实现流程如下：

算法：正向最大匹配法。

输入：一个字符串 S1。

输出：字符串分词后得到的字符数组。

```
begin
1 加载分词词典
2 设置最大分词长度 MaxLen
3 设置输出字符数组 S2 初始为空
4 获取字符串 S1
5 if（S1 不为空）
6      从 S1 左边开始，取候选串 Z，Z 的长度小于 MaxLen
7      if（Z 不在词典中）
8           将 Z 最右边的一个字去掉
9          if（Z 不为单字）
```

```
10              执行 8
11          else
12              S2=S2+Z
13              S1=S1-Z
14      else
15          S2=S2+Z
16          S1=S1-Z
17        执行 6
18 else
19 输出结果 S2
end
```

5.1.2 语句相似度的计算

在本系统中，对中文提问进行分词以后，需要计算中文语句的相似度。代码实现流程如下：

```
算法：加权平均语句相似度计算
输入：两个句子 S1 和 S2。
输出：两个句子的相似度（百分比）。
begin
1 输入两个句子 S1 和 S2
2 foreach 句子 Si（i=1,2）
3       对句子进行分词，生成词向量 W1，W2
4       计算每个词对应在句中的权重，生成权重向量 S1’和 S2’
5 余弦计算相似度 sim=cos（S1’，S2’）
6 返回结果 sim1
7 foreach W1 中所有词
8       foreach W2 中所有词
9           将 W1 中所有词与 W2 中所有词做语义比较，得到一个数组 a
10 foreach W2 中所有词
11      foreach W1 中所有词
12          将 W2 中所有词与 W1 中所有词做语义比较，得到一个数组 b
13 sumA=sumA+a[i]
14 sumB=sumB+b[i]
15 sim2=1/2*( sumA/m+ sumB/n),m,n 分别为数组 a、b 的长度
```

16 sim= α sim1+β sim2
17 返回结果 sim
end

5.2 系统关键模块的实现

5.2.1 登录模块的实现

智能答疑系统的主要用户有学生用户、助教老师用户、名师用户。其中，学生用户输入用户名和密码后需点选“学生”标志，进入后台验证，若验证成功，则成功进入学生用户首页，否则重新进入该登录页面。其他角色登录情况类似，在输入自己的用户名和密码之后需点选正确的角色标志，则能够成功进入该角色信息首页。此外，本系统默认的角色为管理员登录，即不点选任何角色标志情况下，输入正确的管理员登录信息，进入管理员后台知识库管理首页。登录界面如图 5-1 所示：

图 5-1 登录页面

学生用户登陆后，可实现答疑、个人信息修改、个人问题查询、对名师提问的回答及注销登录等功能。学生登录首页如图 5-2 所示：

图 5-2 学生用户登录首页

管理员成功登陆后，可以对知识库进行管理和维护。如：对知识库中问题及答案的分类查询、对知识库中问答信息的添加、对知识库中信息的修改及删除。管理员用户登录首页如图 5-3 所示：

问题编号	问题	回答	访问次数	级别	类别	修改	删除
18	计算机语言的发展过程	机器语言：机器指令的集合，一条机器指令能控制计算机执行一种操作。 汇编语言：用于替代0、1组成的符号指令集合。 高级语言：高级语言编程独立于机器，大大提高编程效率。	23	2	PAR基础		
19	什么是结构程序设计？	机构程序设计是一种进行程序设计的原则和方法，按照这种原则和方法设计出的程序的特点是结构清晰，容易阅读，容易修改，容易验证。	52	2	PAR基础		
20	什么是程序流程图？	流程图是一个描述程序的控制流程和指令的有向图，一个程序可以用流程图的形式表现出来。	33	1	PAR基础		
21	泛型程序设计的作用是什么？	泛型程序设计的作用在于能编写出清晰、易理解、可重用的程序，能反映程序中算法的实质，是编写抽象算法程序的有力工具。	0	0			
22	什么是泛型程序设计？	泛型程序设计是指：在程序设计的过程中，根据需要引入类型或操作作为参数，并以此为基础，编制出具有通用型的程序。	0	0			
23	PAR方法的主要特点是什么？	1. PAR方法是一种统一的设计算法程序的方法。它可以取代已有的分而治之，动态规划，贪心，列举和其他一些不知名的算法设计方法和技术，设计出快速算法，从而避免了在现在各种算法设计方法间做出选择的困难。2.	0	0			
24	统一的算法程序开发方法（PAR方法开发算法程序的步骤）？	1. 用Radl语言精确地描述求解问题的功能规约。2. 把需求解的问题分化成和原问题结构相同但规模更小的子问题，分解可一直进行下去直至每个子问题可直接求解。3. 构造问题求解序列的递推关系Si=F（Sj）并	0	0			
25	什么是抽象程序设计语言Apla？	Apla是我们为实现算法程序形式化开发的PAR方法而定义的一种抽象程序设计语言（Abstract Programming Language）。它是Radl-Apla程序转换器的目标语言，又是Apla到	0	0			

图 5-3 管理员登录首页

5.2.2 智能答疑模块的实现

学生用户登陆后，在首页点选“智能答疑”，进入智能答疑模块。在输入框中输入你想查询的问题，点击“发送”按钮，则系统会及时快速的进行问题获取并与知识库中相同或相似问题进行匹配，并根据相似程度返回最相似问题的答案。在问题的回答后设置学生用户对智能回答模块的评价，若用户选择为对答案满意，则可继续问答；若用户觉得答案不准确或不能理解该回答并选择“不满意”，则系统自动进入助教在线答疑模块。智能答疑模块如图 5-4 所示：

图 5-4 智能答疑页面

此外，对学生的提问，在回答界面的右侧会根据当前问题自动生成系列问题推荐给学生，若有问题恰好为该学生将要提问的问题，点击该问题就可获取该问题答案。智能推荐的设计大大提高了答疑的效率。

5.2.3 助教答疑模块的实现

学生登陆后，可选择进行智能答疑或助教在线答疑。若选择智能答疑且在智能答疑模块中提问并没有得到满意的解答，可进入到助教答疑模块。在本答疑模块中，学生用户输入问题，点击“发送”按钮，完成提问。助教老师登录并查看问题，根据自身掌握的知识对问题进行回答。若学生对助教老师的回答不满意，则可点击“寻求名师帮助”。助教答疑模块如图 5-5 所示：

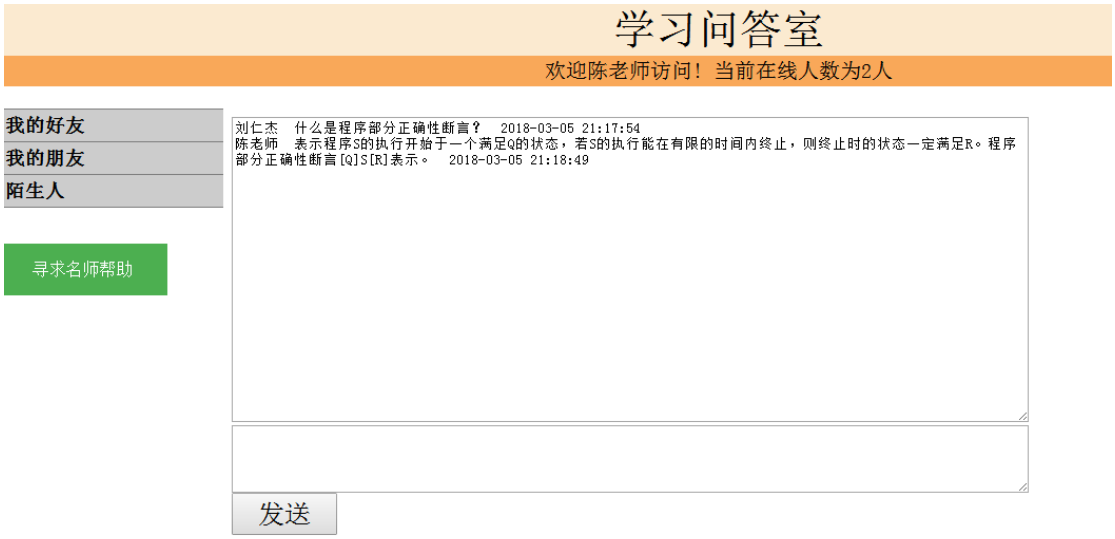

图 5-5 助教答疑页面

远程教育是为了更好地促进教师和学生，学生和学生之间相互交流、讨论而设计的。本模块为学生提供了一个很好的信息共享，交流思想的环境。

5.2.4 名师答疑模块的实现

学生登陆后，可选择进行智能答疑、助教在线答疑或名师答疑。若选择智能答疑或助教在线答疑且问题没有得到满意解答，则可选择名师答疑。由于名师的大量时间都在进行学术研究，并不能及时或全天在线对学生的提问进行解答，系统设计该答疑形式为用户留言形式，即用户在留言板上写下自己的问题并提交，名师在空闲时间对所有用户提交问题进行查看并解答。名师答疑模块如图 5-6 所示：

图 5-6 用户提问页面

5.3 智能答疑系统的应用与分析

本智能答疑系统的设计和实现是在网络教学的环境下产生的，主要用于对程序设计方法学中相关问题的回答。本文请本校中英合作项目组 10 名成员进行了系统的使用与测试，实验的过程和结果可作为衡量本智能系统可靠性和有效性的标准。

5.3.1 权重因子的设定

本文在对语句进行相似度计算时得出最终相似度$M = \alpha T + \beta S$，其中$\alpha + \beta = 1$，产生两个权重因子α、β，为了获取更合适的α、β值使得语句相似度更加准确，本文参看标准问题并提问，获取一组语句及语义相似度数据集。设：$\alpha T + \beta S > 0.85$时，该答案为标准答案。通过一组已知标准答案来进行参数调试，参数计算公式如下：

$$0.85 < \frac{\alpha T + \beta S}{\alpha + \beta} < 1$$
$$\alpha + \beta = 1$$

通过程序不断计算并取得α、β取值范围，最终得出当$\alpha = 0.46$、$\beta = 0.54$时，系统给出的正确答案个数最多，即智能答疑效果最好。

5.3.2 实验过程及结果分析

收集问答对，建立一个测试用例集。随机选取 5 名学生，以自己的方式选取测试集中问题并提问，让系统自动给出答案，分析系统输出结果并将结果与正确答案进行对比，统计系统给出答案正确的个数，计算出智能模块问答的准确率。

（1）中文分词测试

对学生提问领域若干问题的过程中，系统分词效果如表 5-1 所示：

表 5-1 分词结果表

问句	分词结果
什么是结构程序设计语言？	什么\是\结构\程序设计\语言
什么是程序部分正确性断言？	什么\是\程序\部分\正确性断言
什么是程序完全正确性断言？	什么\是\程序\完全\正确性断言
什么是 Floyd 归纳断言法？	什么\是\Floyd 归纳断言法
复杂算法是如何设计的？	复杂\算法\是\如何\设计\的

（2）系统答疑准确率测试

不同的学生登录进行系统测试，测试结果如表 5-2 所示：

表 5-2 查准率测试

学生	提问个数	正确答案个数	查准率	查全率	提问方式
001	20	19	95%	100%	自然语言
002	16	16	100%	100%	自然语言
003	28	27	96%	100%	自然语言
004	25	20	80%	100%	自然语言
005	26	25	96%	100%	自然语言

由表一可以看出，使用本系统智能模块进行《程序设计方法学》知识答疑，平均查准率在 90%以上，查全率为 100%，该模块能够很好的满足用户对概念性知识及公式定理的解答。

（3）系统提问预测效果测试

根据已建立马尔科夫提问预测模型，对该 5 名学生所提问题统计分析，本文设定系统能预测到学生下次提问的知识分类的某一知识点中问题，该问题与学生下次所提问题有 80%的相似度，则该问题为对学生提问问题的准确预测,准确率公式为：

$$\mathrm{pre} = \frac{\text{准确预测个数}}{\text{提问总个数}}$$

统计以上 5 名学生提问及学生下次提问与提问预测的分析比较结果，如表 5-3 所示。不同的学生登录进行系统测试，对每个学生提问，系统给出 5 个下一次提问预测。预测结果统计如表 5-3 所示：

表 5-3 提问预测

学生	提问个数	准确预测提问次数	预测准确率
001	20	14	70.0%
002	16	12	75.0%
003	28	21	75.0%
004	25	17	68.0%
005	26	20	76.9%

由表 5-3 可以看出，使用本系统智能模块进行《程序设计方法学》知识答疑，对学生提问的预测，问题预测平均准确率在 70%以上，可以大大提高系统的答疑效率。

5.4 本章小结

本章主要介绍了智能答疑系统相关技术及功能的具体实现，首先介绍了系统

实现的主要技术，如：中文分词、语句相似度匹配等；然后对系统的主要功能模块进行了展示，包括登录模块、智能答疑模块、助教在线答疑模块、名师答疑模块等；最后设计实验，验证系统的可靠性及答疑的准确性。

6 结　　语

本文论述了智能答疑系统对辅助网络教学的重要性，通过对比国内外研究现状，总结智能答疑系统应有的特性，参考大量的文献资料，提出本系统的研究和设计方案。采用自然语言处理技术，实现了本智能答疑系统的开发；同时，提出隐马尔科夫提问预测模型，并将其首次应用在智能答疑系统中，提高系统的答疑效率。

当前本系统主要应用于对 PAR 方法教学内容的答疑，将问题根据难度划分进行逐层答疑，为老师的教学工作减轻了很大压力。但系统还存在着不足，如：未能对网络资源进行充分利用。本文虽然实现了对知识库的不断扩充和丰富，但是未能将丰富的网络资源库信息利用起来，在后续的工作中考虑在智能模块，将知识库中不能解答的问题进行网络搜索，实现网络资源的共享。

教育是知识创新的基础。智能答疑系统作为网络教学的辅助平台，对提高学生的学习效率有着至关重要的作用。在答疑系统的设计中，智能助教、智能名师等功能的研制，真正实现答疑全程自动化、智能化，也将成为智能答疑系统研究的新目标。

参考文献

[1]王竹立.我国教育信息化的困局与出路——兼论网络教育模式的创新[J].远程教育杂志,2014,32(02):3-12.

[2]王丽莉,孙宝芝.互联网+时代背景下网络教育发展新趋势——“2015 国际远程教育发展论坛”综述[J].中国远程教育,2015(12):12-17.

[3]陈丽.术语“教学交互”的本质及其相关概念的辨析[J].中国远程教育,2004(03):12-16+78-79.

[4]焦建利,贾义敏,任改梅.教育信息化的宏观政策与战略研究[J].远程教育杂志,2014,32(01):25-32.

[5]曾帅,王帅,袁勇,倪晓春,欧阳永基.面向知识自动化的自动问答研究进展[J].自动化学报,2017,43(09):1491-1508.

[6]Figueroa A. Automatically generating effective search queries directly from community question-answering questions for finding related questions[M]. Pergamon Press, Inc. 2017.

[7]郭文俭. 基于课程教学网站的智能答疑系统的设计与实现[D].吉林大学,2015.

[8]张际博. 基于PAR平台与跨媒体的算法程序设计在线教学研究[D].江西师范大学,2017.

[9]熊小舟. 基于Web Service和多媒体数据库技术的PAR方法自学系统的研究[D].江西师范大学,2017.

[10]王东升,王卫民,王石,符建辉,诸峰.面向限定领域问答系统的自然语言理解方法综述[J].计算机科学,2017,44(08):1-8+41.

[11]袁鼎荣, 李新友, 邵延振. 用于中文分词的组合型歧义消解算法[J]. 计算机应用与软件, 2011, 28(6):57-58.

[12]奚宁, 李博渊, 黄书剑,等. 一种适用于机器翻译的汉语分词方法[J]. 中文信息学报, 2012, 26(3):54-58.

[13]苏晨, 张玉洁, 郭振,等. 适用于特定领域机器翻译的汉语分词方法[J]. 中文信息学报, 2013, 27(5):184-190.

[14]王继成, 武港山, 周源远,等. 一种篇章结构指导的中文Web文档自动摘要方法[J]. 计算机研究与发展, 2003, 40(3):398-405.

[15]韩永峰, 许旭阳, 李弼程,等. 基于事件抽取的网络新闻多文档自动摘要

[J]. 中文信息学报, 2012, 26(1):58-66.
[16]解冲锋,李 星.基于序列的文本自动分类算法[J].软件学报,2002(04):783-789.
[17]张俐, 李星, 陆大. 中文网页自动分类新算法[J]. 清华大学学报(自然科学版), 2000, 40(1):39-42.
[18]杨建武. 基于核方法的 XML 文档自动分类[J]. 计算机学报, 2011, 34(2):353-359.
[19]井晓阳, 罗飞, 王亚棋. 汉语语音合成技术综述[J]. 计算机科学, 2012, 39(s3):386-390.
[20]王志明, 陶建华. 文本-视觉语音合成综述[J]. 计算机研究与发展, 2006, 43(1):145-152.
[21]王志明, 蔡莲红, 艾海舟. 基于数据驱动方法的汉语文本-可视语音合成[J]. 软件学报, 2005, 16(6):1054-1063.
[22]冯哲, 孙吉贵, 张长胜,等. 汉语语音合成的研究进展[J]. 吉林大学学报(信息科学版), 2007, 25(2):198-206.
[23]骆卫华, 罗振声, 宫小瑾. 中文文本自动校对技术的研究[J]. 计算机研究与发展, 2004, 41(1):244-249.
[24]刘亮亮, 曹存根. 中文“非多字词错误”自动校对方法研究[J]. 计算机科学, 2016, 43(10):200-205.
[25]刘亮亮, 曹存根. 基于局部上下文特征的组合的中文真词错误自动校对研究[J]. 计算机科学, 2016, 43(12):30-35.
[26]杨淦. 基于条件随机场模型的中文分词系统研究与实现[D].重庆大学,2015.
[27]李月伦,常宝宝.基于最大间隔马尔可夫网模型的汉语分词方法[J].中文信息学报,2010,24(01):8-14.
[28]任惠,林鸿飞,杨志豪.融合字特征的平滑最大熵模型消解交集型歧义[J].中文信息学报,2010,24(04):18-24.
[29]Zhang L Y, Qin M, Zhang X M, et al. A Chinese word segmentation algorithm based on maximum entropy[C].In: International Conference on Machine Learning and Cybernetics. IEEE, 2010:1264-1267.
[30]李春生,卢鹏飞,张可佳.基于语句相似度计算的智能答疑系统机理研究[J/OL]. 计算机技术与发展,2018(03):1-5[2018-0403].http://kns.cnki.net/kcms/detail/61.1450.TP.20171205.0904.026.html.
[31]周永梅,陶红,陈姣姣,张再跃.自动问答系统中的句子相似度算法的研究[J].计算机技术与发展,2012,22(05):75-78.

[32]Rafael Ferreira,Rafael Dueire Lins,Steven J. Simske,Fred Freitas,Marcelo Riss. Assessing sentence similarity through lexical, syntactic and semantic analysis[J]. Computer Speech & Language,2016,39.

[33]庞亮,兰艳艳,徐君,郭嘉丰,万圣贤,程学旗.深度文本匹配综述[J].计算机学报,2017,40(04):985-1003.

[34] Jiwoon Jeon, W. Bruce Croft, and Joon Ho Lee. Finding similar questions in large question and answer archives. In CIKM, pages 84-90, 2005.

[35] Li Cai, Guangyou Zhou, Kang Liu, and Jun Zhao. Learning the latent topics for question retrieval in community qa. In IJCNLP, pages 273-281, 2011.

[36] Zongcheng Ji, Fei Xu, Bin Wang, and Ben He. Question-answer topic model for question retrieval in community question answering. In CIKM, pages 2471-2474, 2012.

[37] Xipeng Qiu and Xuanjing Huang. Convolutional Neural Tensor Network Architecture for Community-based Question Answering. In IJCAI, 2015.

[38] Guangyou Zhou, Tingting He, Jun Zhao and Po Hu. Learning continuous word embedding with metadata for question retrieval in community question answering. In ACL, 2015.

[39] Guangyou Zhou, Yin Zhou, Tingting He and Wensheng Wu. Learning semantic representation with neural networks for community question answering retrieval. In Knowledge-Based Systems, 2015.

[40]荣光辉,黄震华.基于深度学习的问答匹配方法[J].计算机应用,2017,37(10):2861-2865.

[41] Huizhong Duan, Yunbo Cao, Chin-Yew Lin, and Yong Yu. Searching questions by identifying question topic and question focus. In ACL-HLT, pages 156–164, 2008.

[42] Kai Wang, Zhaoyan Ming, and Tat-Seng Chua. A syntactic tree matching approach to finding similar questions in community-based qa services. In SIGIR, pages 187-194, 2009.

[43] Xiaobing Xue, Jiwoon Jeon, and W. Bruce Croft. Retrieval models for question and answer archives. In SIGIR, pages 475-482, 2008.

[44] Guangyou Zhou, Li Cai, Jun Zhao, and Kang Liu. Phrase-based translation model for question retrieval in community question

answer archives. In ACL-HLT, pages 653-662, 2011.

[45] Zhao-YanMing, Tat-Seng Chua, and Gao Cong. Exploring domain-specific term weight in archived question search. In CIKM, pages 1605-1608, 2010.

[46] Li Cai, Guangyou Zhou, Kang Liu, and Jun Zhao. Learning the latent topics for question retrieval in community qa. In IJCNLP, pages 273-281, 2011.

[47] Xin Cao, Gao Cong, Bin Cui, Christian S. Jensen, and Quan Yuan. Approaches to exploring category information for question retrieval in community question-answer archives. ACM TOIS, 30(2): p. 1-38, 2012.

[48] Zongcheng Ji, Fei Xu, and Bin Wang. A category-integrated language model for question retrieval in community question answering. In AIRS, pages 14-25, 2012.

[49]武永亮,赵书良,李长镜,魏娜娣,王子晏.基于 TF-IDF 和余弦相似度的文本分类方法[J].中文信息学报,2017,31(05):138-145.

[50]张齐勋,刘宏志,刘诗祥,贾堂,曹健.基于行业专有词典的 TF-IDF 特征选择算法改进[J].计算机应用与软件,2017,34(07):277-281.

[51]Liu K, Zhang Y Z, Guo-Liang J I, et al. Representation Learning for Question Answering over Knowledge Base:An Overview[J]. Acta Automatica Sinica, 2016.

[52]Liu H, Xu J, Liu C, et al. Semantic-aware Query Processing for Activity Trajectories[C].In: Tenth ACM International Conference on Web Search and Data Mining. ACM, 2017:283-292.

[53]Bagheri E, Bagheri E. Document Retrieval Model Through Semantic Linking[C].In: Tenth ACM International Conference on Web Search and Data Mining. ACM, 2017:181-190.

[54]Zhang Y, Iwaihara M. Evaluating semantic relatedness through categorical and contextual information for entity disambiguation[C].In:Ieee/acis, International Conference on Computer and Information Science. IEEE, 2016:1-6.

[55]Li C, Bendersky M, Garg V, et al. Related Event Discovery[C].In: Tenth ACM International Conference on Web Search and Data Mining. ACM, 2017:355-364.

[56]Arab M, Jahromi M Z, Fakhrahmad S M. A graph-based approach to

word sense disambiguation. An unsupervised method based on semantic relatedness[C].In: Electrical Engineering. IEEE, 2016:250-255.

[57]Xin Y, Xie Z Q, Yang J. Semantic community detection research based on topic probability models[J]. Acta Automatica Sinica, 2015.

[58]秦元巧, 孙国强. 改进的句子相似度计算在问答系统中的应用[J]. 微计算机信息, 2011, 27(8):206-208.

[59]刘宏哲. 一种基于本体的句子相似度计算方法[J]. 计算机科学, 2013, 40(1):251-256.

[60]李玲,何聚厚.基于语义依存分析的句子相似性度量算法及应用研究[J].计算机应用与软件,2017,34(07):244-248+313.

[61]刘雄,张宇,张伟男,刘挺.基于依存句法分析的复合事实型问句分解方法[J].中文信息学报,2017,31(03):140-146.

[62]李冬梅,张琪,王璇,檀稳.基于浅层句法分析和最大熵的问句语义分析[J].计算机科学与探索,2017,11(08):1288-1295.

[63]张仰森,郑佳,李佳媛.一种基于语义关系图的词语语义相关度计算模型[J].自动化学报,2018,44(01):87-98.

[64]朱征宇,孙俊华.改进的基于《知网》的词汇语义相似度计算[J].计算机应用,2013,33(08):2276-2279+2288.

[65]邓涵,朱新华,李奇,彭琦.基于句法结构与修饰词的句子相似度计算[J].计算机工程,2017,43(09):240-244+249.

[66]于根,李晓戈,刘睿,范贤,杜丽萍.基于信息抽取技术的问答系统[J].计算机工程与设计,2017,38(04):1051-1055.

[67]吴国顺. 问题检索与答案排序互相促进的社区问答系统[D].华东师范大学,2017.

[68]李超,柴玉梅,高明磊,咎红英.句法分析和深度神经网络在中文问答系统答案抽取中的研究[J].小型微型计算机系统,2017,38(06):1341-1346.

[69]郑磊,王莉,段跃兴.基于马尔科夫模型的用户兴趣转移建模[J].计算机工程与设计,2018,39(01):177-182.

[70]Slim A, Heileman G L, Kozlick J, et al. Employing Markov Networks on Curriculum Graphs to Predict Student Performance[C].In: International Conference on Machine Learning and Applications. IEEE Computer Society, 2014:415-418.

[71]Hughes G, Dobbins C. The utilization of data analysis techniques in predicting student performance in massive open online courses (MOOCs)[J]. Research & Practice in Technology Enhanced Learning,

2015, 10(1):10.

[72]Homsi M, Lutfi R, Carro R M, et al. A Hidden Markov Model Approach to Predict Students' Actions in an Adaptive and Intelligent Web-Based Educational system[C].In: International Conference on Information and Communication Technologies: From Theory To Applications. IEEE, 2008:1-6.

[73]蒋卓轩, 张岩, 李晓明. 基于 MOOC 数据的学习行为分析与预测[J]. 计算机研究与发展, 2015, 52(3):614-628.

[74]武法提, 牟智佳. 基于学习者个性行为分析的学习结果预测框架设计研究[J]. 中国电化教育, 2016(1):41-48.

致　　谢

岁月如歌，三年的研究生生涯即将结束，我瑾向所有关心、爱护、帮助我的人表示最真诚的感谢和美好的祝愿！

本论文是在钟林辉副教授和薛锦云教授两位老师的悉心指导下完成的，在论文的写作过程中，我遇到了很多困难，两位老师都提供了解决办法，在论文内容上也多次提出修改意见，在此向两位老师表示衷心的感谢。薛老师有着丰富的专业知识、严谨的教学态度，敏锐的学术洞察力。在我的研究生期间给予我很大的帮助，教会我如何总结和凝练知识，如何能够快速高效地学习。三年里，钟老师陪伴我们度过了很多次小组讨论会，教会了我应该从整体的角度判断事物之间的关系。两位老师的教导让我受益终生。

在此，我还要感谢一起愉快地度过研究生生活的国家网络化支撑软件国际科技合作基地的各位师兄师姐以及老师们，最亲爱的游珍师姐、胡启敏师兄、陈媛媛老师、李美玲老师，当然还有我们最可爱的谢武平师兄，感谢你们这三年来对我学习上的帮助和生活上的照顾，祝福你们工作顺心，万事如意！

在读期间公开发表论文（著）及科研情况

攻读学位期间发表论文情况：

[1]Zhong L, Xue L, Zhang N, et al. A tool to support software clustering using the software evolution information[C].In:IEEE International Conference on Software Engineering and Service Science. IEEE, 2017:304-307.

[2] 钟林辉,薛良波,夏鲸等.一种构件化软件演化本体模型的自动构建方法研究[J].计算机应用研究.（已录用）