

学校代码 10126

学号 W1443154

分 类 号

编号

内蒙古大学

工程硕士学位论文

在线答疑系统的设计与实现

研 究 生: 叶 亮

指导教师: 张献国、陈海涛

学 院: 计算机学院

领 域: 计算机技术

2019 年 12 月 9 日

原创性声明

本人声明：所呈交的学位论文是本人在导师的指导下进行的研究工作及取得的研究成果。除本文已经注明引用的内容外，论文中不包含其他人已经发表或撰写过的研究成果，也不包含为获得内蒙古大学及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：

叶亮

指导教师签名：

张永刚

日

期：

2019.12.9

日

期：

2019.12.9

在学期间研究成果使用承诺书

本学位论文作者完全了解学校有关保留、使用学位论文的规定，即：内蒙古大学有权将学位论文的全部内容或部分保留并向国家有关机构、部门送交学位论文的复印件和磁盘，允许编入有关数据库进行检索，也可以采用影印、缩印或其他复制手段保存、汇编学位论文。为保护学院和导师的知识产权，作者在学期间取得的研究成果（含计算机软件、程序）属于内蒙古大学计算机学院。作者今后使用涉及在学期间主要研究内容或研究成果，须征得内蒙古大学计算机学院就读期间导师的同意；若用于发表论文，版权单位必须署名为内蒙古大学计算机学院方可投稿或公开发表。

学位论文作者签名：

叶亮

指导教师签名：

张永刚

日

期：

2019.12.9

日

期：

2019.12.9

在线答疑系统的设计与实现

摘 要

随着计算机网络技术和多媒体技术的快速发展，其相关技术已经被广泛应用到日常教学过程中。本文针对中等职业院校师生设计了在线答疑系统，该系统为学生提供了一种新颖有趣的互动交流方式，为教师提供了方便的答疑途径，能够在调动学生学习积极性、提升教学效果的同时，降低教师的工作强度。本文主要研究工作包括：

（1）本文对在线答疑系统进行需求分析、业务流程分析和数据流程分析，在此基础上对系统进行了整体设计、模块功能设计和数据库设计，然后采用 ASP+MYSQL 框架对系统进行了实现，并对实现结果进行了功能测试。

（2）针对笔者所在的中职院校学生的实际情况，为了调动学生答疑解惑的积极性，系统设计了积分功能，提问和帮助回答问题都会得到相应积分，积分排名和积分兑换功能能够激发学生的学习积极性，进一步促进了答疑系统作用的发挥。

（3）为了减轻教师答疑压力，本文研究了自动问答相关的理论与技术基础，并实现了基于 SimHash 签名的相似问题检索方法，在运行性能和效果上都取得了较好的表现。

关键词：在线答疑；自动问答；信息检索；SimHash

DESIGN AND IMPLEMENTATION OF ONLINE QUESTION ANSWERING SYSTEM

ABSTRACT

With the rapid development of computer network technology and multimedia technology, its related technology has been widely applied to the teaching process. The online question answering system designed for teachers and students in secondary vocational schools provides students with a novel and interesting way of learning and a convenient way for teachers to answer questions. It can mobilize students' learning enthusiasm and improve the teaching effect, while reducing the workload of teachers. The main work of this paper includes:

(1) This paper describes the requirement analysis, business process analysis and data process analysis of the online question answering system. Then, the overall design, module function design and database design of the system are presented. Finally, the system is implemented using ASP+MYSQL framework and the functional test are carried out on the system.

(2) In view of the actual situation of the students in the secondary vocational schools where the author works, in order to mobilize the enthusiasm of students to answer questions and solve puzzles, the reward points function is designed. Asking questions or answering questions will get corresponding reward points. The functions of reward points ranking and reward points exchange can stimulate students' learning enthusiasm and further promote the work of the question answering system.

(3) In order to alleviate the pressure on teachers to answer questions, this paper studies the theoretical and technical basis of automatic question answering, and implements a similar question retrieval method based on SimHash, which achieves a good performance.

KEYWORDS: online question answering; automatic question answering; information retrieval; SimHash

目 录

摘 要.....	I
ABSTRACT.....	II
第一章 绪论.....	1
1.1 研究背景与意义.....	1
1.2 可行性分析.....	2
1.2.1 经济可行性分析.....	2
1.2.2 技术可行性分析.....	2
1.2.3 操作可行性分析.....	3
1.2.4 可行性分析总结.....	3
1.3 国内外研究现状.....	3
1.4 论文的组织结构.....	5
1.5 本章小结.....	6
第二章 系统相关理论和技术基础.....	7
2.1 系统开发工具介绍.....	7
2.1.1 B/S 架构.....	7
2.1.2 ASP 程序设计语言.....	7
2.1.3 MySQL 数据库.....	8
2.2 自动问答相关技术基础.....	8
2.2.1 分词和词性标注.....	9
2.2.2 关键词提取与向量空间表示.....	9
2.2.3 相似性度量.....	10
2.3 本章小结.....	11
第三章 在线答疑系统的分析.....	12
3.1 系统需求分析.....	12
3.2 业务流程分析.....	14
3.2.1 用户登录流程.....	14
3.2.2 问题发布流程.....	15
3.2.3 问题解答流程.....	15
3.3 数据流程分析.....	16
3.3.1 问题和答案数据流图.....	16
3.3.2 积分数据流图.....	17
3.3.3 用户数据流图.....	18
3.4 系统非功能性需求分析.....	18
3.4.1 系统性能.....	18
3.4.2 安全性.....	19
3.4.3 可靠性.....	19
3.5 本章小结.....	19
第四章 在线答疑系统设计.....	20
4.1 系统整体设计.....	20
4.2 模块功能设计.....	22

4.2.1 教师用户管理模块设计	22
4.2.2 学生用户管理模块设计	23
4.2.3 常见问题管理模块设计	24
4.2.4 问题管理模块设计	24
4.2.5 积分管理模块设计	27
4.2.6 相似问题自动检索模块设计	27
4.3 数据库设计	28
4.4 本章小结	32
第五章 在线答疑系统的实现与测试	33
5.1 系统实现	33
5.1.1 用户登录界面	33
5.1.2 登录主界面	33
5.1.3 问题管理模块实现	34
5.1.4 积分管理模块实现	40
5.2 系统测试	41
5.2.1 测试目的	41
5.2.2 系统功能测试	41
5.2.3 系统运行及维护	42
5.3 本章小结	43
第六章 总结与展望	44
6.1 总结	44
6.2 展望	44
参考文献	45
致谢	47

图目录

图 3.1 管理员用例图	12
图 3.2 教师角色用例图	13
图 3.3 学生角色用例图	13
图 3.4 用户登录过程流程图	14
图 3.5 问题发布过程流程图	15
图 3.6 问题解答流程图	16
图 3.7 问题和答案数据流图	17
图 3.8 积分数据流图	17
图 3.9 用户数据流图	18
图 4.1 系统拓扑结构图	20
图 4.2 系统三层架构图	20
图 4.3 系统整体功能架构图	21
图 4.4 添加教师用户流程图	22
图 4.5 编辑教师用户流程图	22
图 4.6 浏览学生用户流程图	23
图 4.7 常见问题的检索和浏览功能流程图	24
图 4.8 学生发布问题流程图	25
图 4.9 回答问题流程图	26
图 4.10 积分兑换流程图	26
图 4.11 系统 E-R 图	29
图 5.1 用户登录页面	33
图 5.2 管理员登录主页面	34
图 5.3 问题管理模块主页面	35
图 5.4 按关键词检索结果页面	35
图 5.5 关键词签名实现的关键代码	36
图 5.6 问句签名实现的关键代码	37
图 5.7 查看问题详情页面	38
图 5.8 问题回答页面	39
图 5.9 问题编辑页面	39
图 5.10 学生积分排名页面	40
图 5.11 积分兑换页面	40

表目录

表 4.1 教师表信息	29
表 4.2 学生表信息	30
表 4.3 问题表信息	30
表 4.4 常见问题表信息	31
表 4.5 待审核问题表信息	31
表 4.6 答案表信息	31
表 5.1 待审核问题模块测试表	41
表 5.2 教师管理模块测试表	41
表 5.3 问题管理模块测试表	42
表 5.4 积分管理模块测试表	42

第一章 绪论

1.1 研究背景与意义

传道授业解惑，是教师工作的几大重要内容，其中，“解惑”对于帮助学生廓清疑难、掌握知识有着重要的作用。而在传统的教学模式中，由于教师的时间限制和对课堂节奏的整齐划一要求，教师往往很难照顾到学生的个性化需求，出现学生问题难以一一解答、解答不够充分等种种问题，最终很可能导致学生积累问题过多、进而对所学内容失去兴趣，这种教与学的矛盾在笔者所在的中等职业院校尤为明显。中等职业院校的学生与普通高中学生相比，学习积极性相对较低，学习动机相对更为匮乏，不少学生虽然也有主动求知的意愿，但往往由于基础知识薄弱，学习能力不足，往往更容易体验到学习中的挫败感。这就需要中职教师在教学中更加关注到每个学生的需求，对学生学习中出现的问题进行及时解答和及时帮助。但是由于学生人数的众多和集体教学的特点，这种要求显然很难实现。

随着互联网和人工智能技术的飞速发展和教育信息化建设的开展，基于网络的教学和问答逐渐成为教师解决教学问题的一种辅助手段，而中职学生恰恰对信息技术等新兴事物很感兴趣，愿意尝试新的学习工具，接受新的学习途径。如果利用信息技术辅助课堂教学，采用网络答疑与传统教学相结合的教学模式，不仅能够解决以上对学生答疑解惑不够充分及时的问题，还能够增加教师与学生之间的互动，甚而增强学生与学生之间的交互。利用网络答疑系统，为学生营造一个全方位、多层次的学习空间，引导学生随时随地通过网络向老师和同学们求教和讨论，从而将被动式学习转变为主动探索。

基于以上想法，笔者针对所教课程的特点以及教学过程中遇到的问题，决定设计与实现一个基于网络的在线答疑系统，希望通过此系统解决教师在集体授课过程中遇到的答疑不及时的问题。该系统拟以学生为核心，以教师为根基，以智能化方法为工具。一方面，学生通过本系统可随时提出自己学习中存在的疑问，获得系统或老师、同学的帮助，从而及时解决问题，为后续学习奠定基础。另一方面，系统通过建立知识库来收集和积累学生的问题，对于知识库中已有的相似问题，系统可以通过相似问题搜索自动推荐问题和答案，减轻教师重复答疑的工作量。对于系统知识库中没有的新问题，再发送

给教师和学生进行讨论和答疑，并以教师作为答案的最终确认者，在教师本人给出问题的正确答案或判定学生给出的答案准确后，将相关问题和答案收集进系统知识库中。随着系统知识库的不断累积和丰富，大部分的学生问题都可以通过相似问题搜索获得解决，从而将教师从重复工作中解放出来，将更多的精力放在教学创新上。最后，教师作为答案的确认者，在建立知识库的过程中，既实现了无纸化远程答疑，及时地满足了学生个性化需求，也辅助教师在一定程度上获知学生在学习过程中产生的问题，进而及时调整教学策略，改变授课方式，以期达到良好的教学效果。在这个过程中，学生的提问和老师对应的解答会慢慢积累成一定可重复利用的网络教学资源，在一定程度上满足广大中职师生的教学和学习需求。

1.2 可行性分析

可行性研究的意义在于，在最初的设计阶段用较小的代价发现错误的构思以及相应的问题能否被解决，从而避免时间、人力、资金等资源投入的浪费。可行性研究的结果也可以将开发风险和系统收益控制在可接受的范围内。为此，在系统设计和开发之初，笔者从经济、技术和操作三个方面进行了系统开发的可行性分析。

1.2.1 经济可行性分析

笔者所在的中职学校能为本在线答疑系统所需的计算机设备和相关软硬件提供经费保障。本系统的服务宗旨是为提高学生的学习积极主动性，方便快捷的促成师生间的学习交流，并非想获得经济上的收益，且系统的开发维护成本也较低，目标用户在使用过程中仅会产生网络流量，基本不会有经济负担，所以此系统在经济上是可行的。

1.2.2 技术可行性分析

基于目前的网络环境，本在线答疑系统的前台应用程序开发工具选用 ASP，后台数据库建立工具选用 MySQL，前后台的连接通过 ASP 中的 ADO 控件实现，而且笔者学校现有的计算机设备和软件环境也完全能够满足开发的需要，可以实现在线答疑功能的设计。

1.2.3 操作可行性分析

就目前笔者所在学校的信息化建设情况看，学校的软硬件的集成化程度很高，操作也很简便，同时本系统在开发过程中充分听取了广大师生的需求和建议，也充分考虑到了用户操作的方便，系统界面简单，操作便捷。

1.2.4 可行性分析总结

本答疑系统的开发成本较小，系统维护也仅需少量人力，主要的师生用户使用本系统只会产生网络流量，当前的网络环境下，这几乎不会对用户造成经济负担。笔者所在学校的信息化建设水平较高，能为本系统开发所需的硬件设备和软件环境提供支持和保障。系统的开发充分考虑了目标用户的特点，系统设计界面清爽大方，操作简单便捷。所以，本在线答疑系统在经济、技术和操作性上都是可行的。

1.3 国内外研究现状

近年来，国内对在线答疑系统研究较多。目前的在线答疑主要包括两种答疑方式：一种是人工答疑，另一种是通过访问数据库来实现自动答疑。

人工答疑的形式主要有电子邮件、论坛等方式。论坛是比较普遍的一种答疑形式，老师与学生、同学之间可针对某个问题展开讨论，这种方式一方面可以促进学生对知识的深化理解、培养他们的沟通学习能力，另一方面教师的参与和激励可激发学生学习的积极性和主动性。电子邮件私密性好，适合个性化学习。以上人工答疑的方式在答疑解惑方面发挥了重要的作用，但人工答疑往往会有较大延时，即学生在提出问题后，教师大部分情况下无法同时在线，一定程度上会降低学生学习的积极性。另一方面，对于相同或相似的问题，教师重复回答也会增加教师的工作量。所以人工答疑的实时性不强、教师工作量大。为了实现网络教学的高效答疑，建立快捷有效的自动答疑系统就显得很有必要。

自动答疑系统以存储量非常大的数据库作为后台支撑，学生在该系统上以自然语言形式提问，系统自动对这个问题进行分析，在抽取其中的关键字、语句理解的基础上，进行知识抽取和知识推理，然后根据检索算法在数据库中进行检索，将检索到的结果合成为答案并返回给系统，这种自动答疑的方式显然比人工答疑时效性更强。自动答疑技

术起源于上世纪五十年代, Protosynthes^[1]是早期典型的答疑系统, 该系统根据文本中的非停用词的词频和依存分析来搜索获得最终答案。到了上世纪八十年代, 出现了以Lunar^[2]为代表的问答系统, Lunar 是用来对月球岩石和土壤数据进行查询和分析的专家系统, 该时期出现了大量针对特定领域的专家系统, 以便于专业领域人员能够快速检索有用信息。MURAX^[3]是上世纪九十年代典型的答疑系统, 它把百科全书作为搜索基础, 来应对各种用户提出的问题。MURAX 能从语法角度分析问题中单词之间的关系, 进而从百科全书中搜索出相应答案。国内早期的在线答疑系统代表是上海交通大学设计开发的Answer Web^[4]自动答疑系统, 该系统的问题和答案数据库是动态化的, 即如果输入的关键词在数据库中存在, 系统就返回相应答案; 如果不存在, 系统则会转给人工帮助给出答案。同时, 系统会将新的问题答案对添加到相应数据库中。

近些年, 随着计算机技术的飞速发展和广泛应用, 答疑系统的智能化程度也越来越高。浙江师范大学的方利伟^[5]开发了一个针对不同用户开展个性化服务的答疑系统, 该系统可以根据学习者的特征, 主动为学习者推送其存在的疑问或感兴趣的知识, 提高系统答疑解惑的能力。中国海洋大学的张银^[6]发现已有的在线答疑系统的开发设计走进了偏重技术的误区, 其造成的现状就是答疑系统的智能化程度越来越高, 提问者却仍难得到满意的解答。针对这一现状, 张银从网络答疑系统的学习机制入手, 通过探究答疑的本质, 提出答疑系统的设计新思想和实现策略。

在答疑系统的关键技术实现方面, 相关研究也有很大推进。例如, 华南理工大学的蒋昌金^[7]等人提出了一种组合词识别算法, 该算法能识别网页上多数的新词和未登录词, 分词效果得到很大的改进; 华东师范大学的朱甜甜^[8]提出了一个针对句子层面的基于多样化特征融合的短文本相似度量模型, 有效的提升了短文本语义相似度量模型的性能; 中国科学技术大学的伍浩铖^[9]通过挖掘用户搜索意图的方法来处理基于查询短关键词的问题搜索排序问题, 该方法可以显著提升短关键词上问题搜索的性能; 江苏大学的陈伟鹤^[10]等人提出了基于词或词组长度和频数的关键词提取算法, 在面向中文短文本的关键词提取中, 达到了更高的准确性; 电子科技大学的张帅帅^[11]通过阅读理解模型来提高答案匹配问题的精度。最近几年, 在问答系统设计与实现方面出现了较多的工作, 南京邮电大学的曹艳蓉^[12]设计与实现了基于中文社区的智能问答系统, 南京大学的张苏可^[13]介绍了悟空问答社区系统服务端的设计与实现, 北京邮电大学的桑志杰^[14]对生成式问答系统技术研究实现进行了详细介绍, 重庆师范大学的申豪杰^[15]对基于知识图谱的电影知识问答系统进行了研究实现, 北京邮电大学的刘珮^[16]针对阅读理解任务设计与实现了

一个问答系统。

随着深度学习技术的发展,许多研究者将深度学习技术应用到在线答疑系统中。具有代表性的产品包括苹果公司研发的 Siri^[17]、微软公司研发的 Cortana^[18] 和小冰^[19]、百度研发的小度^[20] 等。在科研方面, Bordes 等^[21] 首次将神经网络中的词嵌入方法^[22] 应用到知识库问答领域。Yhi 等^[23] 则将卷积神经网络模型应用到了答案推理中。Dong^[24] 也采用了卷积神经网络来对答案进行更好的理解。基于神经网络的方法不需要对自然语言进行人工解析,不需要人工设计规则和特征,对自然语言处理工具的依赖性较低,适用于大规模的自动问答应用系统,但神经网络方法的不足之处是需要大规模的训练数据。

以上的自动答疑系统的关键技术层出不穷,智能化程度也越来越高,但是,许多关键技术和算法并不公开,大型的系统也很难应用到中职教学答疑这样的“小问题”解决层面上。但是,笔者遇到的中职教学中的答疑困难确实存在,一个可应用推广的面向中职课程教学的答疑系统的实现也确实有其实践意义和应用价值。作为工作在基层一线的专业教师,笔者仍然觉得在借鉴现有成熟技术和算法的基础上,设计开发一个可以解决一线教学实际问题的在线答疑系统,具有实际价值。

1.4 论文的组织结构

根据研究工作的内容,本文通过六个章节,详细地介绍了在线问答系统的设计与实现,具体安排如下:

第一章绪论,主要介绍本研究的研究背景与意义、可行性分析以及国内外研究现状。

第二章系统相关的理论与技术基础,主要介绍 Web 设计技术基础和系统相关智能化算法的理论基础。

第三章在线答疑系统的分析,主要涉及系统需求分析、业务流程分析和数据流程分析等。

第四章在线答疑系统的详细设计,主要包括系统整体设计、模块功能设计和数据库设计等。

第五章在线答疑系统的实现与测试,主要包括系统主要界面展示、系统主要功能的测试结果等。

第六章总结与展望,对系统实现过程和结果进行总结以及对未来工作进行展望。

1.5 本章小结

本章首先介绍了研究背景与意义，然后介绍了可行性分析与研究现状，最后给出本文结构，为后文撰写奠定基础。

第二章 系统相关理论和技术基础

本章首先介绍了系统设计和开发过程中用到的相关技术，然后介绍了系统后台用到的自动问答的相关理论基础。

2.1 系统开发工具介绍

2.1.1 B/S 架构

B/S（浏览器/服务器）架构是一种被广泛应用于广域网上的结构模式，Web 浏览器是客户端最主要的应用，这种模式将系统功能实现的核心部分放在服务器上，非核心部分放到客户端上。B/S 架构的优势包括：

1. 升级和维护方式简单：相比于 C/S（客户端/服务器）架构，B/S 架构的系统升级过程简单方便，系统管理人员只需要对服务器端升级维护即可，而不用担心用户规模带来的影响，这种以瘦客户端、胖服务器端的架构，会极大的减少人力、物力、时间等资源的投入。

2. 跨平台：现今浏览器已经成为各个操作系统的标准配置，使得 B/S 架构具有很好的跨平台特性。系统在服务器端集中管理控制，服务器无需知道客户端浏览器运行的系统平台，而且浏览器也不会知道服务器的系统信息，因此安全性更高。

3. 客户端负荷轻：由于系统的业务逻辑主体都放在了服务器端实现，只有很少的控制浏览器端实现，网络管理员只负责管理服务器，基本实现了客户端零维护。

2.1.2 ASP 程序设计语言

ASP（动态服务器页面）是微软开发的一种简单实用的软件开发编程语言，它能与目前流行的数据库和其它软件开发程序语言实现交互功能。ASP 技术主要有以下几个特点：

1. ASP 语言无需编译。ASP 编写的脚本集成于 HTML 当中，易于修改和测试，由 Web 服务器的引擎进行解释并执行，无需编译或链接即可直接解释执行。

2. ASP 是纯文本格式的程序文件。ASP 不需要特定的开发编程工具，使用常规文本

编辑器(如 WINDOWS 下的记事本),即可进行程序代码的开发,保存代码时只需将文件扩展名改为“.ASP”即可。

3. ASP 独立于浏览器。ASP 程序是通过所在的服务器引擎来解释执行的,用户端的浏览器只需接受并读懂服务器端所返回的用户请求即可。目前绝大多数浏览器都支持用户阅读使用 ASP 语言开发的网页功能。

4. ASP 的高兼容性。ASP 程序语言能与其他脚本语言实现高度的兼容性,这是在 ASP 设计之初就被考虑的特点,因此也赢得了广泛用户的支持。

2.1.3 MySQL 数据库

MySQL 是目前流行的关系数据库系统之一, MySQL 在与其它程序语言的接口方面,因其独特的灵活性而深受好评。MySQL 数据库操作管理系统体积小、速度快且是免费开源的,它为开发人员提供了所需的大多数功能。同时,MySQL 具有良好的可移植性,能在多数时下流行的操作系统中运行,易于使用。

MySQL 数据库操作管理系统自带的优点和功能总结如下:

1. 自身管理功能,MySQL 数据库操作管理系统在用户环境下的并发控制、对不同用户的使用权限的设置、数据库的运行日志记录和管理以及数据库的安全检查等方面具有优异的管理性能。

2. 数据管理功能,MySQL 允许用户对数据库内的数据进行分类、查询、添加、删除、修改、更新等操作,其数据管理效率和管理方法在同类型的数据库中更胜一筹。

3. 数据的可扩展性,MySQL 虽在设计之初被定位成一个小数据库应用系统,但在允许的范围内具有很好的可扩展性,这点对于一些实时交互式站点系统进行实时数据更新非常重要,而且 MySQL 也可以最大程度的实现数据管理系统对数据可扩展性管理的自动化,大大提高了数据库系统的应用范围。

4. 可靠的数据保护功能,MySQL 在数据加载的同时,会进行数据合法性检测。如果出现数据删除、丢失、保存失败等情况,MySQL 可以进行数据恢复,具有实时数据备份的功能。

2.2 自动问答相关技术基础

自动问答的常见形式包括“常见问题”回答系统和社区自动问答等形式,“常见问题”

回答系统中通过专业技术人员搜集和整理问题及其对应答案，然后发布在系统中供客户浏览并找到相应的解决方案。而社区问答系统往往比较复杂，一方面问题冗长而多样，并且可能会对问题进行追加解释，另外关于问题还存在语境、起因等影响因素。另一方面，问题的答案没有专业人员对其逐一整理给出。这就导致社区自动问答系统的实现面临诸多挑战。本系统借鉴了“常见问题”回答系统和社区自动问答系统的各自优点，针对重点难点知识点，以“常见问题”形式罗列出来，以便学生及时解决遇到的知识点问题。针对平时练习过程中遇到的问题，以社区问答的形式实现，方便师生间的探讨。该系统用到的关键技术包括分词和词性标注、特征词提取与表示、问句相似性度量方法等。

2.2.1 分词和词性标注

中文的句子不像英文句子由空格分开每个单词，因此中文分词是中文信息处理的基本步骤，实现中文分词的算法主要包括基于词典的方法和基于机器学习的方法^[25-27]，基于词典的方法是按照一定的方法将待匹配的字符串与词典中的单词进行匹配，如果找到对应单词则识别成功。常见的匹配方法包括正向匹配法、逆向匹配法、最大匹配法和最小匹配法等。基于机器学习的分词方法是通过构建词频语料库来判断一个字符串是否构成词语，相邻的字共同出现的次数越多，构成一个词的可能性就越大，主要统计模型包括 N-gram（N 元语法模型）、HMM（隐马尔可夫模型）等。

在分词之后需要将停用词去除，停用词本身不具有实际意义，去除之后能够减少数据冗余，提高问题句子处理的精度。可以通过定义一个停用词表来去除停用词，例如，“的”、“呀”等。分词之后还有一个重要工作就是词性的标注。如果判断出一个词属于何种词性，如名词、动词、形容词、限定性副词等，就能判断该词在上下文中的作用。本文采用 `cppjieba`^[28] 中文分词工具完成分词工作。

2.2.2 关键词提取与向量空间表示

在自动问答系统中，提取问句的关键词^[29]是描述问句的基础，关键词是描述问句的核心，目前抽取关键词的方法主要分为有监督和无监督两大类，有监督方法是需要提前标注训练集，然后通过训练集上训练一个关键词抽取器，以此来提取问句的关键词。无监督的方法是通过某种方法来实现对候选词按照重要性从高到低进行排序，选择排名靠前的候选词作为关键词，目前常用的无监督关键词提取方法包括基于词频统计、基于

潜在语义以及基于图模型的关键词提取方法等。

下面介绍基于图模型的 TextRank^[30] 算法来提取问句的关键词。TextRank 算法来源于 Google 的 PageRank 算法，TextRank 算法的步骤大致为：首先对问句进行分词、去除停用词，然后以单词为结点，通过 N-Gram 思路，将单词和它附近的 N 个单词对应的结点都连一条无向边，这样就建立了无向图。然后将每个结点 V_i 的初始权重（重要程度） $S(V_i)$ 设为 1，阻尼系数 d 设置为 0.85。最后根据以下公式迭代计算出每个单词结点的权重：

$$S(V_i) = (1 - d) + d \sum_{j \in \text{In}(V_i)} \frac{1}{|\text{Out}(V_j)|} S(V_j) \quad (1)$$

其中 $\text{In}(V_i)$ 是指能够连接到结点 V_i 的那些结点， $|\text{Out}(V_j)|$ 是指 V_j 能够连接到的结点的数目。

在得到问句的关键词之后就可以采用 BOW（词袋）模型将问句映射到特征向量空间。BOW 模型的基本思路为：首先建立词典，其中包含 W 个单词。然后根据词典中的词是否在问句中出现将向量对应位置设为 0 或 1，或者根据单词的重要程度设置为浮点数值，这样针对每个问句 q 可以建立其向量空间表示为：

$$q = (x_0, x_1, \dots, x_W) \quad (2)$$

其中， x_i 如果为大于 0 的值，则表明第 i 个单词在问句中出现，如果为 0 则表明问句中没有出现该单词。

2.2.3 相似性度量

在自然语言处理领域，相似性度量技术被广泛应用于文本摘要、信息检索、文本分类、机器翻译等应用。在得到问句的向量空间表示之后，可以根据向量之间的距离来描述问题之间的相似程度，常用的距离度量方法有点积、余弦距离、欧式距离、马氏距离等。除此之外，常用的汉语句子相似度计算方法还包括基于句法特征的方法和 SimHash^[31] 方法等。

SimHash 方法是 Google 为了解决传统哈希方法无法给出语句相似度的问题而形成的，SimHash 形成的两个二进制签名，其海明距离能够反映两个语句的相似程度，距离越小相似程度越高。除此之外，SimHash 方法的搜索效率也比较高。该方法具体步骤如下：

1. 分词：将问句分词，并求关键词及其权重。例如对于语句：“在线答疑系统的设计与实现”，分词后为：“在线，答疑，系统，的，设计，与，实现”，然后提取关键词及权重：在线(4)，答疑(5)，系统(3)，设计(5)，实现(5)，其中，括号中的数字越大，代表单词在句中的重要性越高。

2. 哈希：利用传统哈希函数将关键词映射成 m 位的二进制签名。例如将上面例子中的关键词映射为：在线(100100)，答疑(010110)，系统(111001)，设计(011011)，实现(101010)，其中括号里的数字代表这个单词的二进制签名。

3. 加权：将第 2 步形成的二进制签名的各个位乘以权重系数，二进制位的值为 1 则用 1 和权重相乘，否则用 -1 和权重相乘，例如将上面例子中的签名加权结果为：在线(4 -4 -4 4 -4 -4)，答疑(-5 5 -5 5 5 -5)，系统(3 3 3 -3 -3 3)，设计(-5 5 5 -5 5 5)，实现(5 -5 5 -5 5 -5)，其中括号里的数字代表这个单词的二进制签名加权结果。

4. 求和：将各个关键词的第 3 步计算结果对应位相加。例如上例相加结果为：

$$(4-5+3-5+5, -4+5+3+5-5, -4-5+3+5+5, 4+5-3-5-5, -4+5-3+5+5, -4-5+3+5-5) = (2 \ 4 \ 4 \ -4 \ 8 \ -6)$$

5. 二值化：第 4 步计算结果各个位的值和 0 比较，大于 0 则置为 1，否则置为 0，最终得到一个 m 位的二进制 SimHash 值作为问句的签名。据此可得上例 SimHash 签名为(111010)。

在得到问句的签名之后，就可以计算问句和数据库中的候选问题签名之间的海明距离，从而可以检索得到前 k 个最相关的候选问题。

2.3 本章小结

本章对系统的开发工具和后台相关的关键算法基础进行了简单介绍，为后文的需求分析、系统设计与实现的介绍奠定基础。

第三章 在线答疑系统的分析

3.1 系统需求分析

软件的需求分析是指用户对目标系统在功能、性能、行为、设计约束等方面的期望。通过对要解决的问题及其应用环境的理解，纠正需求模糊性，排除用户的不合理需求，挖掘潜在需求，在此基础上制定需求说明文档，以此指导开发人员进行软件系统的设计和实现。

本系统用户角色包括管理员、教师和学生三种。管理员作为系统必不可少的一部分，其角色主要功能包括系统维护、数据库管理和维护、教师管理、学生管理、内容审核。系统初始化时拥有一个默认管理员 admin。管理员用例图如图 3.1 所示。

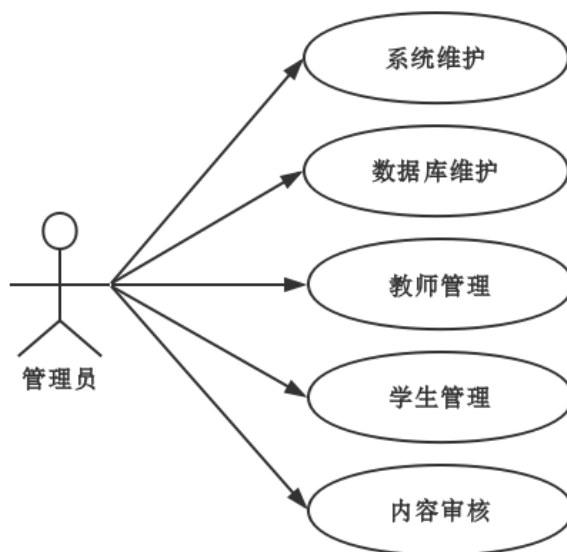


图 3.1 管理员用例图

Figure 3.1 Administrator's use case

教师作为学生学习的引导者，需要解答学生的疑问或者判断学生的答案，需要设计相应的问题引导学生思考，还要根据重难点知识整理学生易错问题及答案，以及根据学生的积分排名判断学生的积极性。因此，在本系统中的教师角色主要功能包括问题解答或答案确认、发布抢答问题、整理常见问题及答案、查看学生积分排名等，教师用户由管理员分配用户名和密码。教师角色用例图如图 3.2 所示。

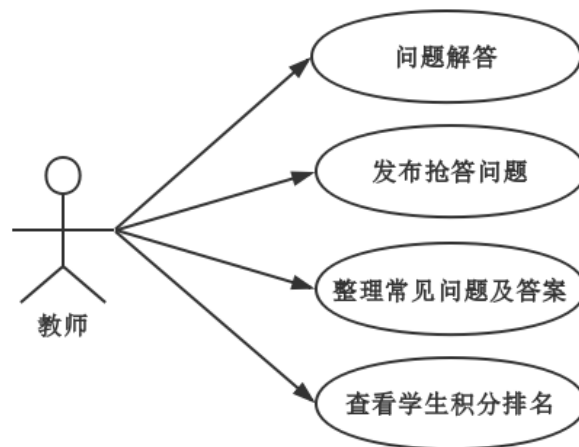


图 3.2 教师角色用例图

Figure 3.2 Teacher's use case

学生在学习中产生疑问时，一方面希望得到老师和同学的帮助，另一方面也希望通过搜索查看相似问题及答案找到解决问题的思路。此外，还要通过查看老师整理的常见问题及答案来巩固知识点的掌握，以及通过查看自己在同学中的积分排名来激励自己力争上游。因此，本系统的学生角色功能包括发布问题、查看和解答别人的问题、问题搜索、查看常见问题及答案、查看积分排名等。学生角色用例图如图 3.3 所示。

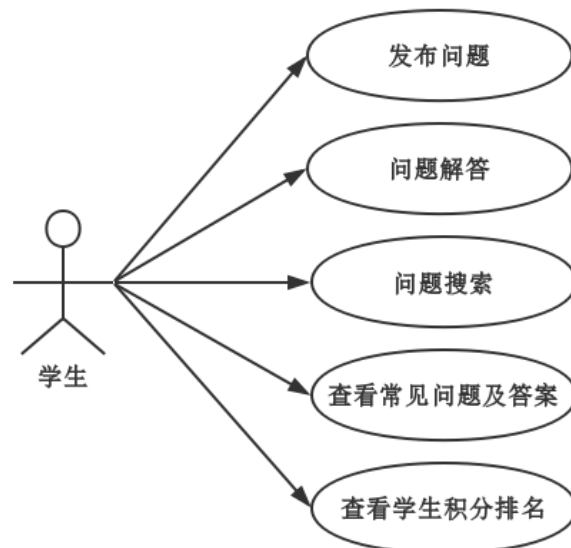


图 3.3 学生角色用例图

Figure 3.3 Student's use case

3.2 业务流程分析

3.2.1 用户登录流程

用户登录过程流程如图 3.4 所示。在用户登录成功后，系统需要根据用户角色显示不同的功能模块。

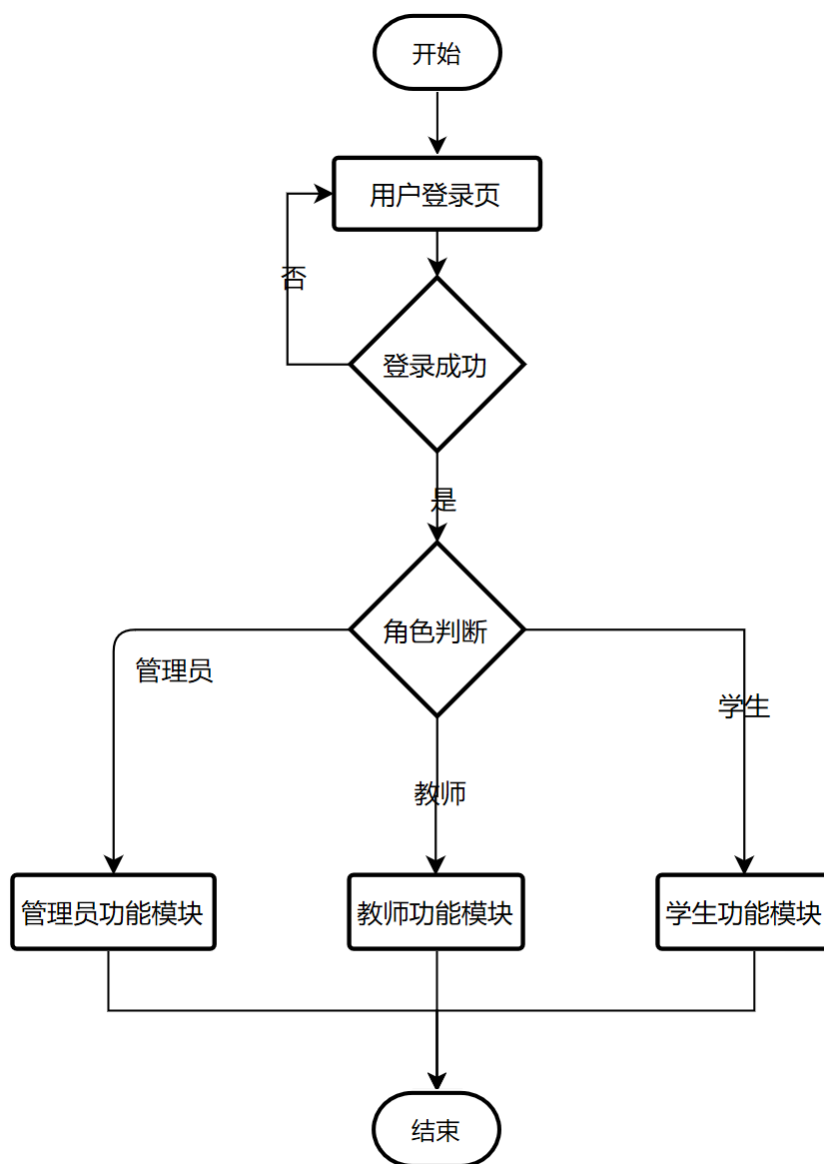


图 3.4 用户登录过程流程图
Figure 3.4 Diagram of the user login process

3.2.2 问题发布流程

问题发布过程流程如图 3.5 所示，学生选择相应科目发布问题，发布的问题需要由管理员审核通过后才能成功发布。教师发布的问题则无需管理员审核直接发布成功。

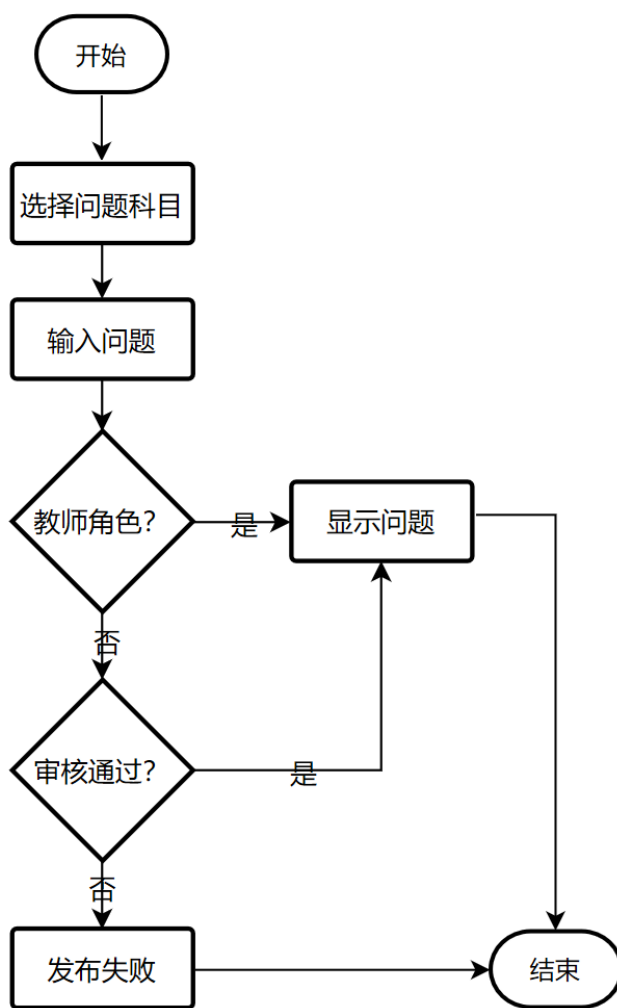


图 3.5 问题发布过程流程图

Figure 3.5 Diagram of the problem release process

3.2.3 问题解答流程

问题解答流程图如图 3.6 所示。学生在回答问题时需要经过教师确认答案是否正确，正确的回答为正确答案状态，错误的回答会被标记为错误答案状态。正确答案和

错误答案及其状态都会被显示出来，以便学生借鉴。教师的回答则直接显示为正确答案状态。

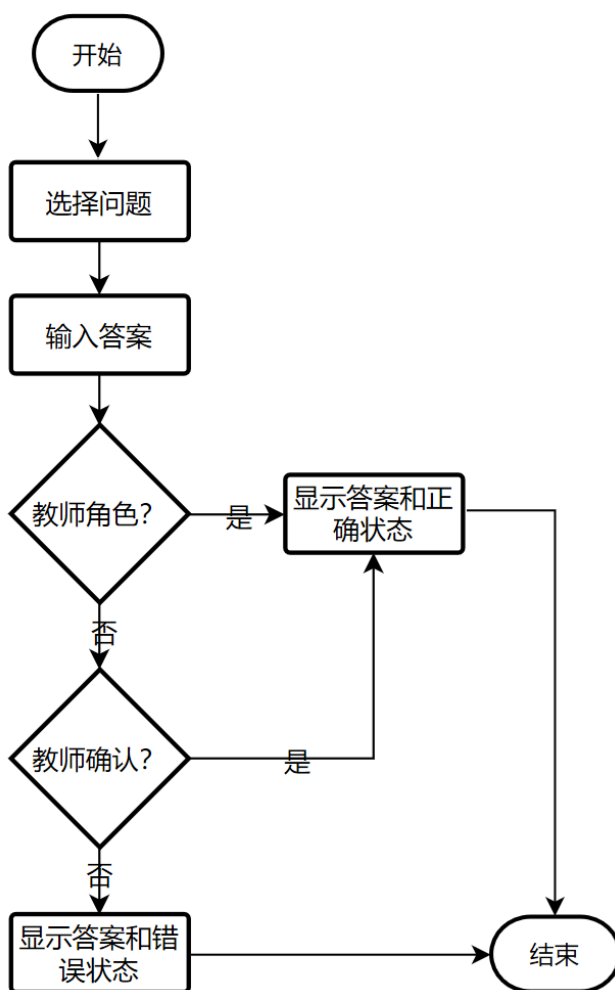


图 3.6 问题解答流程图

Figure 3.6 Diagram of the problem answering

3.3 数据流程分析

3.3.1 问题和答案数据流图

问题和答案数据流图如图 3.7 所示，教师和学生可以发布问题和解答问题，也可以编辑或删除自己发布的问题，但管理员不可进行上述操作，管理员负责审核学生发布的问题，教师负责审核学生的解答。

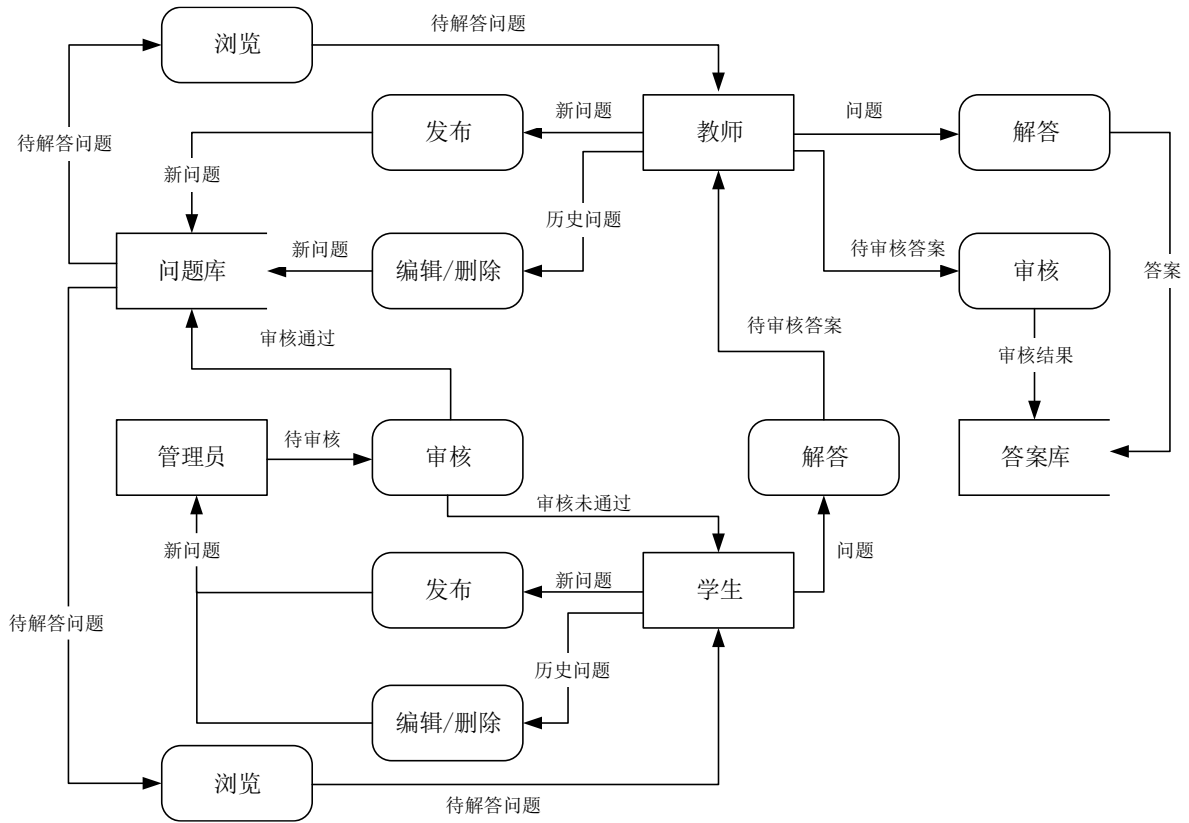


图 3.7 问题和答案数据流图

Figure 3.7 Data flow diagram about questions and answers

3.3.2 积分数据流图

积分数据流图如图 3.8 所示，学生通过发布问题、成功解答问题获得积分，通过积分兑换功能会扣除相应积分。

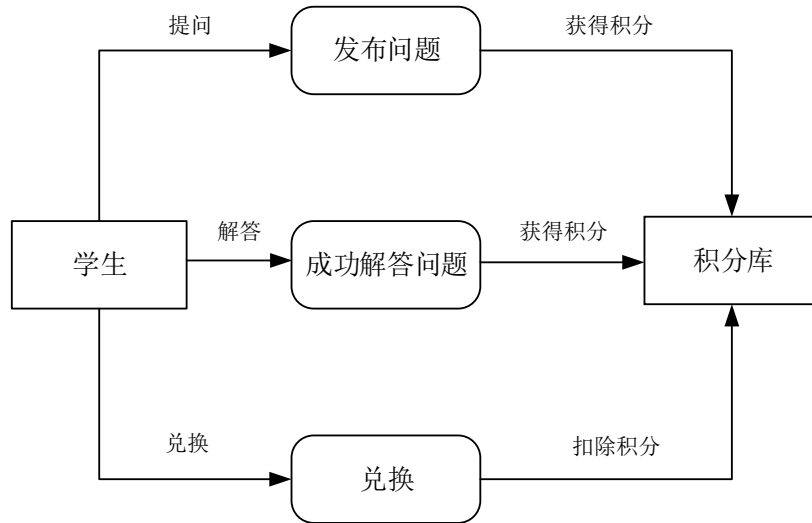


图 3.8 积分数据流图

Figure 3.8 Data flow diagram about user points

3.3.3 用户数据流图

用户数据流图如图 3.9 所示，用户数据通过管理员申请添加，个人仅拥有修改和编辑个人信息的权限，用户角色通过用户 ID 字段来判断。

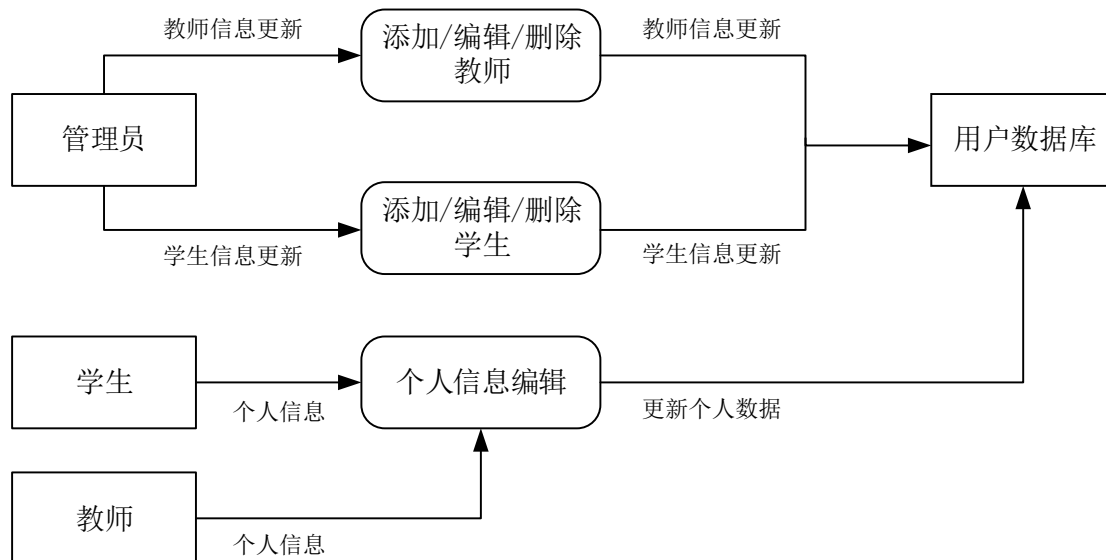


图 3.9 用户数据流图

Figure 3.9 Data flow diagram about user information

3.4 系统非功能性需求分析

非功能性需求是需求分析的一个重要组成部分，为了满足系统所需要实现的目标，需要对非功能性需求进行详细分析。

3.4.1 系统性能

系统性能包括用户能接受的响应时间、系统能够容纳的数据规模等。根据本系统应用场景，用户数约为 1 万人，每天登录用户数为 3000 人左右，网络的带宽为 100M 带宽。系统需要同时满足 1 万个用户请求，要求在 95% 的情况下，一般时段响应时间不超过 2 秒，高峰时段不超过 5 秒。在非高峰时间，根据关键词或问句进行搜索，可以在 3 秒内得到搜索结果。系统容量需要支持 5 万用户，支持 TB 级数据。数据库表行数可以超过 100 万行。CPU 占用率小于 50%，内存占用率小于 50%。

3.4.2 安全性

不同的用户应具有不同的身份和权限，只有经过身份认证后的用户，才能访问其权限范围内的数据，才能操作其权限范围内的各项功能。这样才能保护用户数据不被非法或越权访问和篡改，确保数据的完整性和安全性。要求系统设置防火墙，能经受来自互联网的一般性恶意攻击，如病毒、木马攻击、猜测攻击、黑客入侵等，并要求至少 99% 的攻击需要在 10 秒内检测到。网络传递数据应经过加密，保证数据在收集、传输和处理过程中不被窃取和篡改。业务数据需要在存储时进行加密，确保不可破解。

3.4.3 可靠性

对需要用户输入的地方进行提示，对输入的数据要有相应检查，以防止数据输入异常。要求系统具有很好的健壮性，能处理系统运行过程中出现的各种异常情况，如人为操作错误、输入非法数据等。因软件系统的失效而造成不能完成业务的概率要小于 5‰。要求系统 7x24 小时运行，全年持续运行故障停运时间累计不能超过 24 小时。系统缺陷率每月最多发生 1 次故障，每学期最多出现 1 次需要重新启动系统的情况。

3.5 本章小结

本章进行了系统需求分析、业务流程分析、数据流程分析以及系统非功能性需求分析，理清了用户需求和系统功能模块间的关系，为详细设计和实现奠定基础。

第四章 在线答疑系统设计

4.1 系统整体设计

系统的拓扑结构图如图 4.1 所示，通过 Web 服务器，使得系统的数据存储服务器和客户端进行了隔离，并且用户需要通过验证才能访问系统，一定程度上保证了数据的安全。

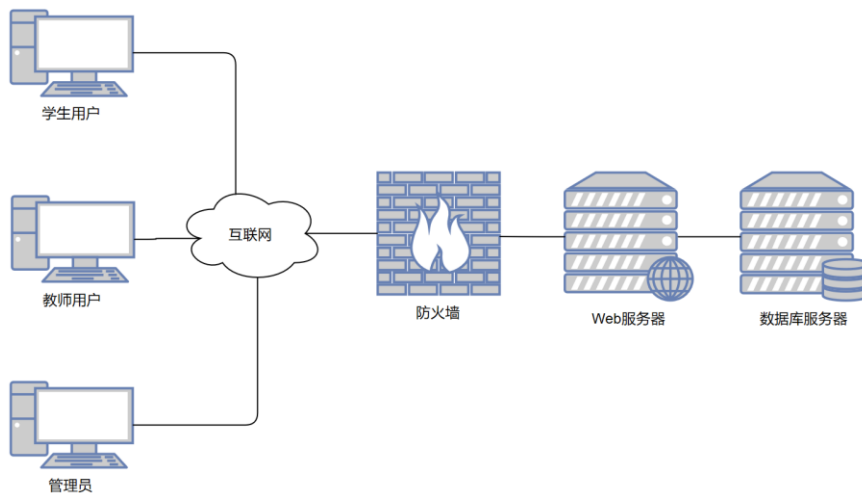


图 4.1 系统拓扑结构图

Figure 4.1 Topological structure of the system

系统体系结构采用 B/S 架构模型，分别以表现层、业务逻辑层、数据访问层展现，层之间相互独立，分工明确，体现了高内聚、低耦合的程序设计思想。对数据统一有效的传输和管理，避免了数据流通时的操作冲突，实现了数据流通的安全性。系统总体架构图如图 4.2 所示。

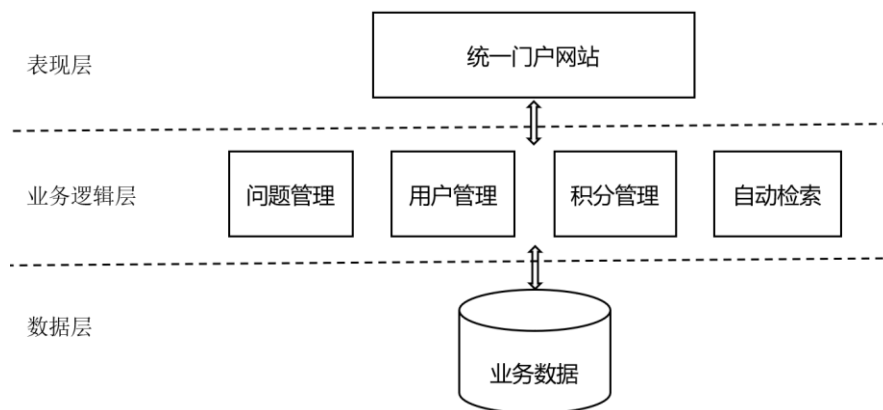


图 4.2 系统三层架构图

Figure 4.2 3-tier architecture of the system

系统整体功能架构图如图 4.3 所示。系统功能包括教师管理、学生管理、问题管理、常见问题管理和积分管理等模块，教师管理和学生管理模块仅管理员可访问，管理员可以通过这两个模块管理教师和学生信息，另外管理员需要对学生发布的问题进行审核，这是管理员在问题管理模块特有的功能。相比于学生用户，在常见问题管理模块中，教师用户拥有问题的发布、修改、检索和浏览功能，而学生仅拥有检索和浏览功能。针对问题的解答，如果是学生给出的答案，需要教师审核通过后方可作为正确答案。针对问题的检索功能，学生提出问题后，系统后台会根据相似问题检索算法给出相似问题，如果学生发现问题一致或对解题思路有帮助，即可点击查看该问题的解答，很大程度上能够快速解决自己的问题，另外，学生也可以通过关键词搜索相关问题及其解答，找到解决问题的思路。在浏览问题时，可以根据更新时间排序浏览，也可以按照科目过滤后再浏览相关问题。为了激发学生解决问题的积极性，系统会给出做出贡献的学生奖励积分，例如发布问题奖励 1 分，正确回答问题可获得 10 积分。积分可以兑换学校提供的小礼品或兑换平时分等，而且学生可以看到每个人的每学期的积分排名，激励学生努力学习、力争上游。

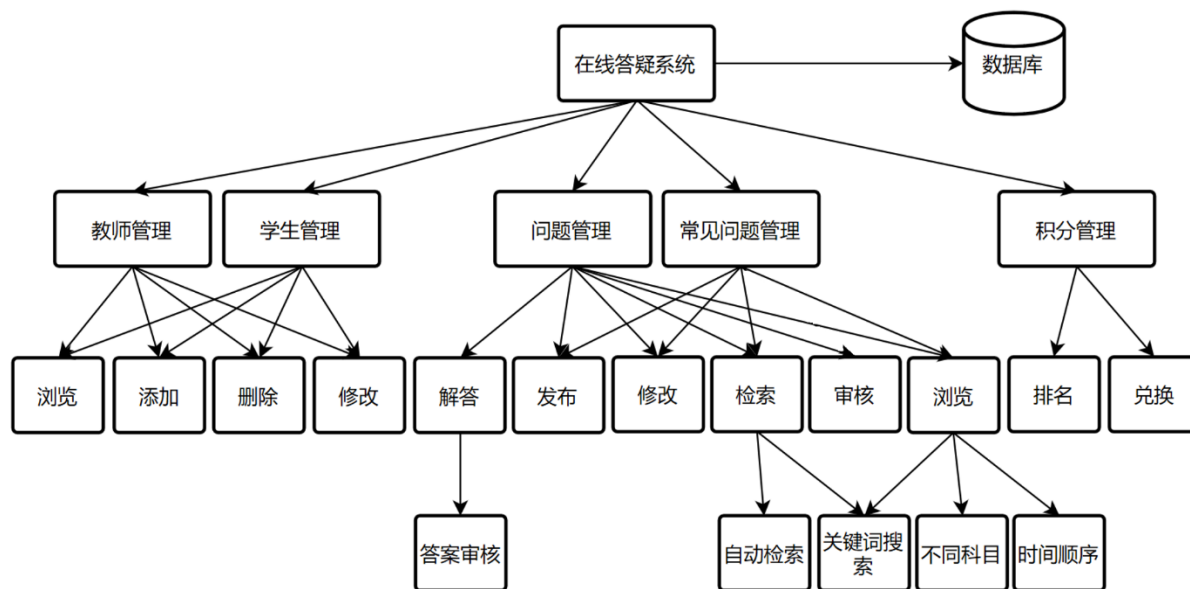


图 4.3 系统整体功能架构图

Figure 4.3 Architecture diagram of the overall system function

4.2 模块功能设计

4.2.1 教师用户管理模块设计

教师用户管理模块功能包括添加教师、编辑教师、删除教师、浏览教师信息。管理员添加教师用户流程图如图 4.4 所示。

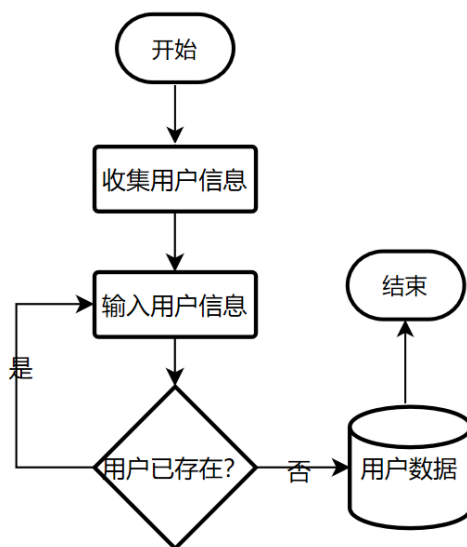


图 4.4 添加教师用户流程图

Figure 4.4 Flowchart of the adding teacher user

编辑教师用户流程图如图 4.5 所示。

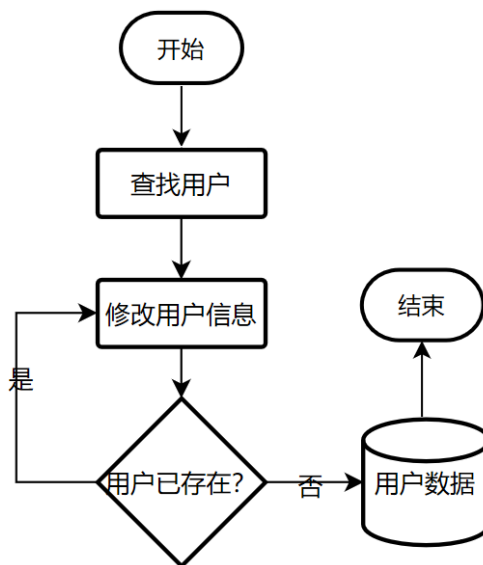


图 4.5 编辑教师用户流程图

Figure 4.5 Flowchart of the editing teacher user

删除教师用户流程和编辑教师用户流程类似，在查找到用户之后点击删除即可。浏览教师用户功能是按照教师工号从小到大排序展示，管理员可以根据工号或姓名查找相应教师用户信息。

4.2.2 学生用户管理模块设计

学生用户管理模块功能包括添加学生、编辑学生、删除学生、浏览学生信息。其中添加、编辑、删除学生用户和添加、编辑、删除教师用户流程类似。在浏览学生用户信息时，本系统提供了按照班级浏览学生和按照学号或姓名查找学生的功能。具体浏览学生用户流程如图 4.6 所示。

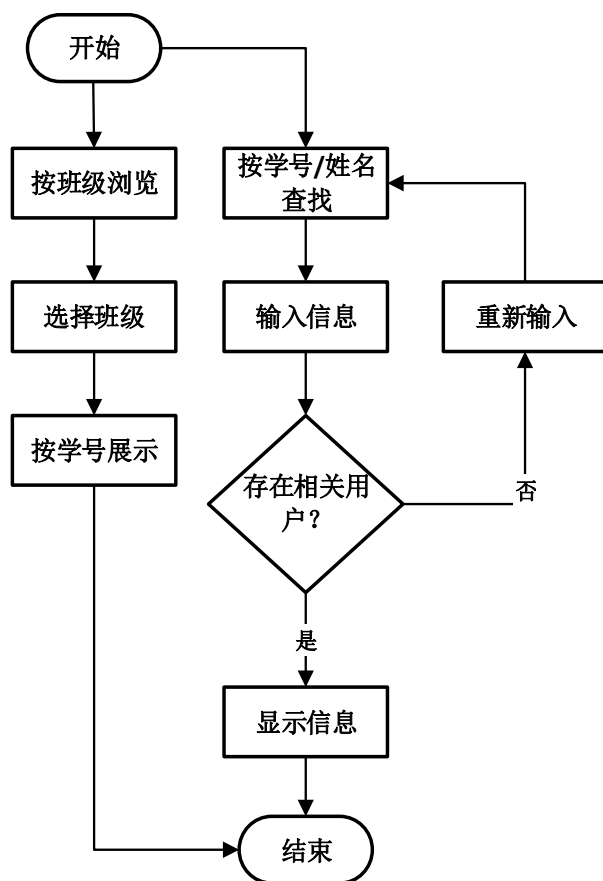


图 4.6 浏览学生用户流程图

Figure 4.6 Flowchart of the browsing student user

4.2.3 常见问题管理模块设计

常见问题管理模块主要包括问题与答案的发布、修改、检索以及浏览功能。常见问题与答案的发布功能仅教师用户拥有该权限，教师整理好重、难点问题对应的常见问题和答案后，将相关信息输入到系统即可完成发布。问题与答案的修改功能也仅教师用户拥有权限，教师用户在查找到需要修改的问题后点击编辑按钮，输入修改信息即可完成修改。教师和学生用户都拥有常见问题的浏览和检索功能。用户在浏览常见问题时，可以根据更新时间排序浏览，也可以按照科目过滤后再浏览相关问题。针对检索功能，系统为用户提供关键词搜索和按问句检索两种方式。常见问题的检索和浏览功能的实现流程如图 4.7 所示。

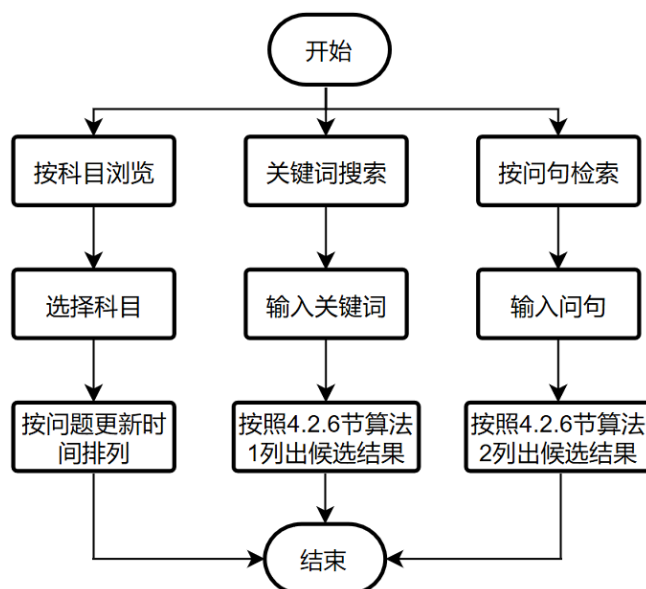


图 4.7 常见问题的检索和浏览功能流程图

Figure 4.7 Flowchart of the search and browse for common problems

4.2.4 问题管理模块设计

常见问题是由教师针对学生常见的问题而整理并添加的，其作用是方便学生快速解决学习过程中遇到的问题。而问题管理模块的功能是为了方便学生提出自己难以解决的问题，希望得到同学或者老师的答复，该模块能够增强师生之间、同学之间的互动，增强学习兴趣。学生提问的流程如图 4.8 所示。学生提问需要由管理员审核通过后方能发布成功，教师也可以发布问题让学生抢答，教师发布问题不需要管理员审核

即可发布。问题修改、删除、检索和浏览功能的实现和常见问题的对应功能的实现类似。

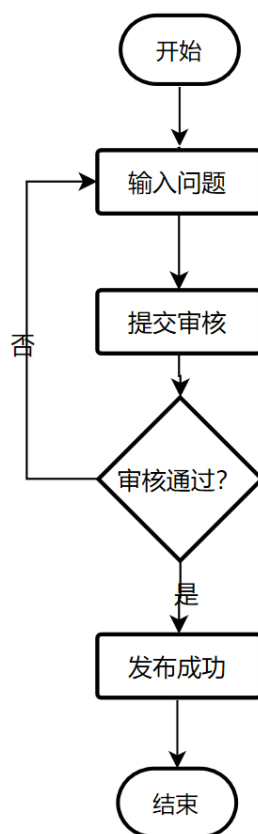


图 4.8 学生发布问题流程图
Figure 4.8 Flowchart of the Student issue

问题管理模块中问题的回答流程如图 4.9 所示。在学生回答问题时，需要教师审核，教师审核为正确的答案则被标记为正确答案，学生会获得相应积分，错误的答案也会显示在页面上，以便于其他学生借鉴。老师的答案则不需要审核，直接显示为正确答案。值得注意的是，一个问题的正确答案可能有多个，这样也有利于学生开拓思路。答案在显示的时候先显示正确的答案，再显示其他答案。

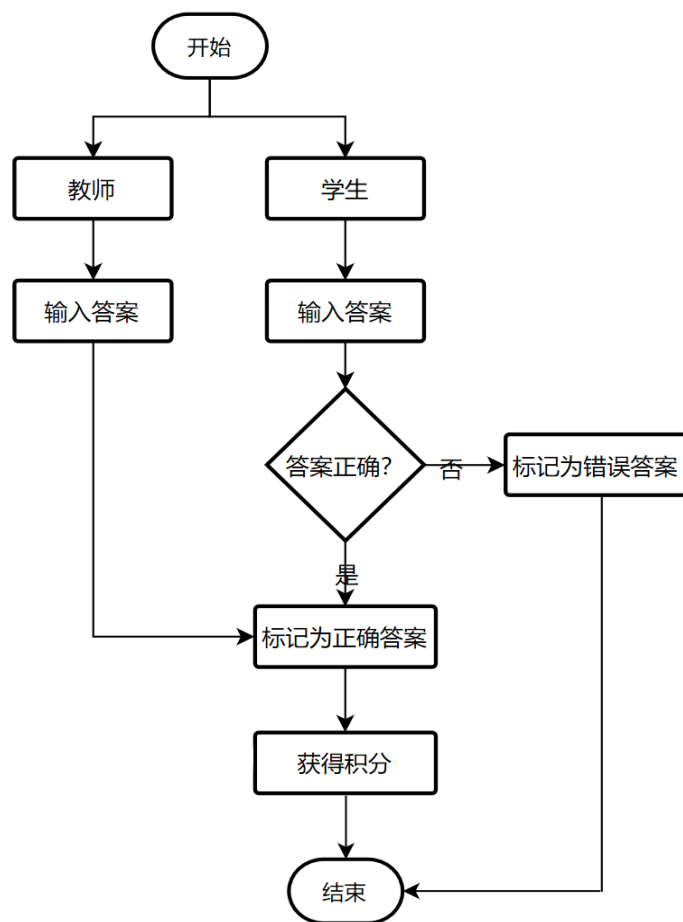


图 4.9 回答问题流程图

Figure 4.9 Flowchart of the answering question

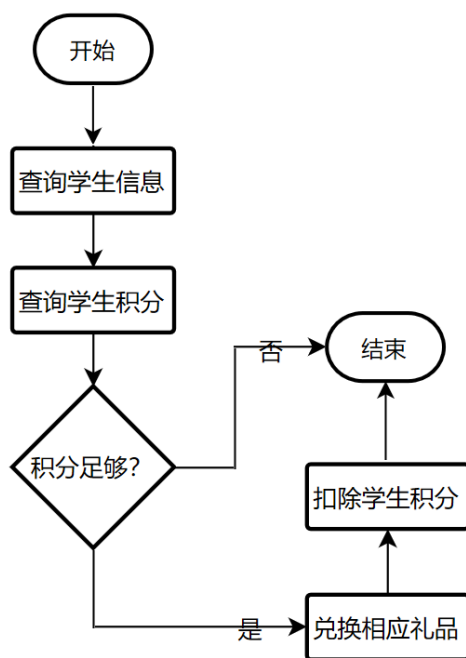


图 4.10 积分兑换流程图

Figure 4.10 Flowchart of the redeeming points

4.2.5 积分管理模块设计

积分管理模块的功能包括积分排名显示和积分兑换功能，积分兑换功能由管理员负责扣除积分和兑换奖品，兑换流程如图 4.10 所示。

积分排名功能包括班级内排名和全校排名两种方式，兑换的积分仍然算在总积分内参与排名，即以当前积分和已兑换积分的总分当作排名积分，排序算法采用的是数据库内置的排序算法，在数据查询的时候按照总分排序获得排序后的结果，然后再显示相关信息。

4.2.6 相似问题自动检索模块设计

在常用问题管理模块和问题管理模块中用到了相似问题的自动检索功能，包括根据关键字检索相似问题和根据问题检索相似问题两种。根据关键字检索问题的算法如算法 1 所示，在该算法中， K 表示关键词的集合， k_i 表示第 i 个关键词， T 表示关键词的个数，哈希函数采用 64 位 Jenkins 哈希函数^[32]。由于用户输入的关键词无法计算每个关键词的权重，所以本系统将所有关键词同等对待，利用类似于 SimHash 的算法计算多个关键词的签名。在根据问题检索相似问题时，根据算法 2 检索相似问题。在添加问题时，系统需要根据 SimHash 算法计算问题对应的签名并保存在数据库中，以便算法 1 和算法 2 使用。

算法 1：基于关键词的相似问题检索算法

输入：关键词 $K=\{k_1, \dots, k_T\}$

输出：候选问题按相似程度的排序结果

- (1) 利用哈希函数将 $\{k_1, \dots, k_T\}$ 关键词均映射成 m 位的二进制签名 $\{h_1, \dots, h_T\}$ ；
 - (2) 对于每个 h_i ，将其各个位拉伸到 -1 和 1，二进制位的值为 1 则不变，值为 0 则变成 -1，拉伸结果记为 s_i 例如，二进制签名为 1010，则拉伸结果为 1, -1, 1, -1；
 - (3) 计算所有关键词对应的 s_1 到 s_T 的各个位对应累加和 S ；
 - (4) S 的各个位的值和 0 比较，大于 0 则置为 1，否则置为 0，最终得到一个 64 位的二进制值作为所有关键词的签名 a ；
 - (5) 计算签名 a 和问题库的每个签名之间的海明距离；
 - (6) 根据海明距离排序结果，返回相似问题检索结果。
-

这个算法的关键在于计算多个关键词的 SimHash 编码, 假设用户输入的关键词为 {“函数”, “图像”, “导数”}, 首先, 通过哈希函数将每个关键词映射成一个 64 位的二进制签名, 此处以 4 位为例进行解释说明。假设关键词“函数”的哈希结果 h_1 为“1011”, “图像”的哈希结果 h_2 为“0101”, “导数”的哈希结果 h_3 为“1100”; 然后, 根据第 (2) 步将 h_1 、 h_2 、 h_3 的哈希结果进行拉伸, 得到 s_1 为“1, -1, 1, 1”, s_2 为“-1, 1, -1, 1”, s_3 为“1, 1, -1, -1”; 接着计算累加和 $s_1+s_2+s_3$ 为“1-1+1, -1+1+1, 1-1-1, 1+1-1”, 即 S = “1, 1, -1, 1”; 最后根据 S 的各个位的值和 0 之间的关系, 确定所有关键词的签名 a 为“1101”。

在计算查询的哈希签名和数据库中问题的签名之间的距离时, 采用海明距离进行计算, 签名 a 和 b 对应位上编码不同的位数即为 a 和 b 之间的海明距离, 可以通过对 a 和 b 进行按位异或运算后统计异或结果是 1 的位数来计算。

算法 2: 基于问句的相似问题检索算法

输入: 问句 q

输出: 候选问题按相似程度的排序结果

- (1) 按照 2.2.3 节介绍的 SimHash 算法计算问句 q 的签名 a , 其中哈希函数采用 Jenkins 哈希函数;
 - (2) 计算签名 a 和问题库的每个签名之间的海明距离;
 - (3) 根据海明距离排序结果, 返回相似问题检索结果。
-

算法 2 首先采用分词工具对问句进行分词和关键词提取, 然后通过类似于算法 1 的步骤来计算关键词的 SimHash 编码, 不同的是在对 h_i 拉伸之后需要再乘以关键词对应的权重, 具体见 2.2.3 节算法介绍。

4.3 数据库设计

本系统的数据库设计 E-R 图如图 4.11 所示, 用到的数据表包括教师表(Teacher)、学生表(Student)、问题表(Question)、常见问题表(Common)、待审核问题表(Waiting)、答案表(Answering)。

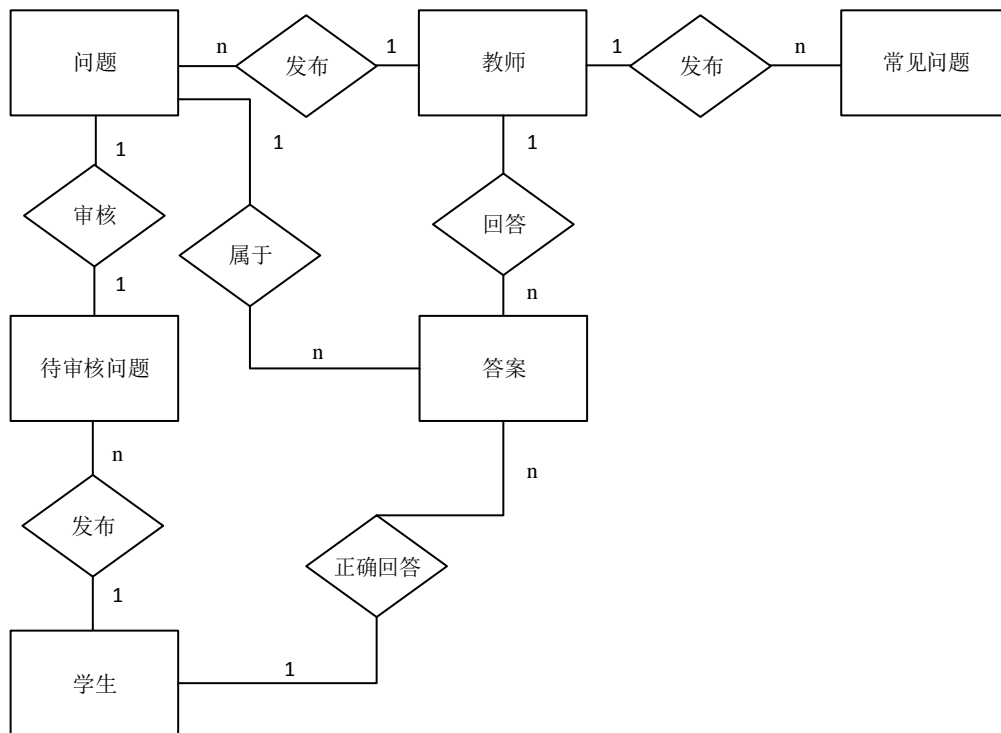


图 4.11 系统 E-R 图

Figure 4.11 E-R diagram of the system

教师表如表 4.1 所示，教师的工号是主键和登录用户名，在分配账号时赋予每个教师一个默认密码，发布问题、回答问题、审核问题的数量都置为 0。

表 4.1 教师表信息

Table 4.1 Information of the Teacher table

列名	类型	是否主键	是否索引	可否为空	说明
UserID	Varchar(20)	Y	Y	N	教工号，作为登录用户名
UserName	Varchar(20)	N	N	N	教师姓名
PassWd	Varchar(30)	N	N	N	登录密码
Num_q	Int	N	N	N	发布问题数量
Num_a	Int	N	N	N	回答问题数量
Num_c	Int	N	N	N	审核回答数量

学生表如表 4.2 所示，也是以学号为主键和登录用户名并分配默认密码，学生发布问题和回答问题的数量、积分、已用积分都置为 0，学生的其他信息由管理员从教务系统的数据库中导出使用。

表 4.2 学生表信息

Table 4.2 Information of the Student table

列名	类型	是否主键	是否索引	可否为空	说明
UserID	Varchar(20)	Y	Y	N	学号, 作为登录用户名
UserName	Varchar(20)	N	N	N	学生姓名
PassWd	Varchar(30)	N	N	N	登录密码
Num_q	Int	N	N	N	发布问题数量
Num_a	Int	N	N	N	回答问题数量
Class	Varchar(20)	N	N	N	班级
Grade	Varchar(20)	N	N	N	年级
College	Varchar(20)	N	N	N	学院
Points	Int	N	N	N	积分
UsedPoints	Int	N	N	N	已用积分

问题表如表 4.3 所示, 问题的 ID 是自增的整数, 提问者的 ID 是学生的学号或教师的工号, 被回答次数能够一定程度上反映该问题的难度, 正确回答日期是第一次被正确回答的日期, SimHash 编码是在问题审核通过后由系统自动计算并填入数据库。问题的答案是存放在答案表中的, 通过问题 ID 关联。

表 4.3 问题表信息

Table 4.3 Information of the Question table

列名	类型	是否主键	是否索引	可否为空	说明
ID	Int	Y	Y	N	自增
Description	Varchar(1000)	N	N	N	问题描述
Question_by	Varchar(20)	N	N	N	提问者 ID
Is_answered	Boolean	N	N	N	是否被正确回答
Q_Date	Datetime	N	N	N	提问日期
A_Date	Datetime	N	N	N	正确回答日期
Num_a	int	N	N	N	被回答次数
Type	Varchar(10)	N	N	N	科目
Num_v	Int	N	N	N	被浏览次数
SH_Code	Int	N	N	N	SimHash 编码

常见问题表如表 4.4 所示, 常见问题和答案是预先由教师整理好的, 仅供学生浏览, 没有回答的相关信息, 答案直接存放在该表中。

表 4.4 常见问题表信息

Table 4.4 Information of the Common table

列名	类型	是否 主键	是否 索引	可否 为空	说明
ID	Int	Y	Y	N	自增
Description	Varchar(1000)	N	N	N	问题描述
Question_by	Varchar(20)	N	N	N	发布者 ID
Answer	Varchar(1000)	N	N	N	答案
Type	Varchar(10)	N	N	N	科目
Num_v	Int	N	N	N	被浏览次数
SH_Code	Int	N	N	N	SimHash 编码
Q_Date	Datetime	N	N	N	发布日期

待审核问题表如表 4.5 所示，待审核问题一旦被审核则信息会从该表删除。

表 4.5 待审核问题表信息

Table 4.5 Information of the Waiting table

列名	类型	是否 主键	是否 索引	可否 为空	说明
ID	Int	Y	Y	N	自增
Description	Varchar(1000)	N	N	N	问题描述
Question_by	Varchar(20)	N	N	N	提问者 ID
Q_Date	Datetime	N	N	N	提问日期
Type	Varchar(10)	N	N	N	科目

答案表如表 4.6 所示，问题的 ID 用作关联答案和问题。

表 4.6 答案表信息

Table 4.6 Information of the Answering table

列名	类型	是否 主键	是否 索引	可否 为空	说明
ID	Int	Y	Y	N	自增
Q_ID	Int	N	N	N	问题 ID
Description	Varchar(1000)	N	N	N	答案描述
Answer_by	Varchar(20)	N	N	N	回答者 ID
State	Int	N	N	N	0: 未审核, 1: 正确答案, 2: 错误答案
A_Date	Datetime	N	N	N	回答时间

4.4 本章小结

本章首先进行了系统整体设计，在此基础上对系统各个模块和数据库进行了详细设计，为后续实现和测试奠定基础。

第五章 在线答疑系统的实现与测试

5.1 系统实现

根据设计开发的需要和实际拥有的条件，本研究开发语言选用 ASP，服务器操作系统为 Microsoft Server 2016，系统后台数据库选用 MySQL 8.0.1，WEB 服务管理器选为 IIS。

5.1.1 用户登录界面

用户登录界面如图 5.1 所示，用户输入用户名和密码后点击登录，系统会判断用户名是否为“admin”，如果是则为管理员用户，否则从教师表和学生表查找对应的用户名 ID 进行判断。

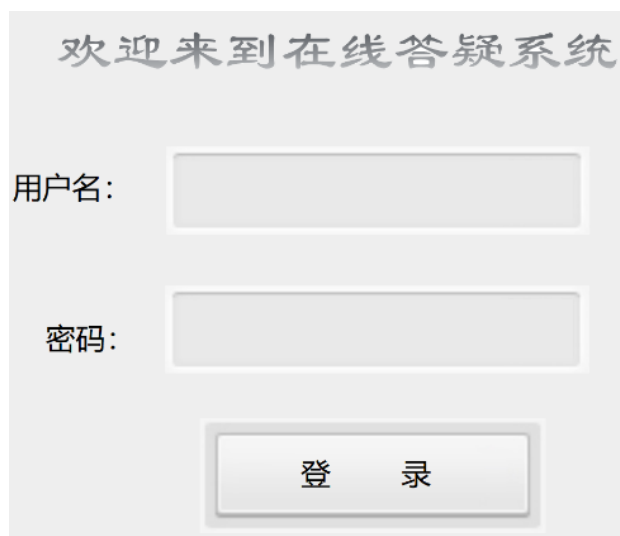
The image shows a user login interface with a light gray background. At the top, the title '欢迎来到在线答疑系统' (Welcome to the Online Q&A System) is displayed in a dark gray font. Below the title, there are two input fields: the first is labeled '用户名:' (Username:) and the second is labeled '密码:' (Password:). Both fields are empty and have a light gray border. Below the password field, there is a button labeled '登 录' (Login) in a dark gray font, centered within a light gray rectangular button.

图 5.1 用户登录页面

Figure 5.1 Illustration of the user login page

5.1.2 登录主界面

管理员用户登录成功后页面如图 5.2 所示。首先显示的是待审核问题，以便于管理员及时审核。教师和学生用户登录后的页面仅包含问题管理、常见问题管理模块，见 5.1.3 节具体实现。



图 5.2 管理员登录主页面

Figure 5.2 Illustration of the administrator login page

在主页面上, 需要从问题表中检索出未审核状态的问题以及根据发布人 ID 找到发布人的姓名信息以便于管理员查看。教师管理和学生管理页面的实现结果和主页类似, 不同的是下面表格部分展示的是教师和学生的相关信息, 可以编辑和删除教师和学生信息。

5.1.3 问题管理模块实现

图 5.3 是问题管理模块主页面, 主页包含根据科目、关键词检索和根据问句检索问题的功能, 罗列了问题的问题描述、发布者、科目、发布时间和是否回答的状态。按照关键词“函数”搜索结果如图 5.4 所示。按照关键词检索和按照问题检索的关键在于计算关键词和问题的签名, 相关代码如图 5.5 和图 5.6 所示, 按照问题检索结果展示页面和图 5.4 类似。



图 5.3 问题管理模块主页面

Figure 5.3 Illustration of the problem management module page



图 5.4 按关键词检索结果页面

Figure 5.4 Illustration of the keyword-based searching page

```

bool keywordHash(vector<pair<string, double> > wordweights, vector<pair<uint64_t, double> >& res)
{
    res.resize(wordweights.size());
    for(size_t i = 0; i < res.size(); i++)
    {
        res[i].first= jenkins_hasher(wordweights[i].first.c_str(), wordweights[i].first.size(), 0);
        res[i].second = wordweights[i].second;
    }
    return true;
}

bool getKeywordSimHash(vector<pair<string, double> > wordweights, size_t topN, uint64_t& retval)
{
    vector<pair<uint64_t, double> > hashvalues;
    if(!keywordHash(wordweights, topN, hashvalues))
    { return false; }
    vector<double> weights(LEN, 0.0);
    const uint64_t val1(1);
    for(size_t i = 0; i < hashvalues.size(); i++)
    {
        for(size_t j = 0; j < LEN; j++)
        {
            weights[j] += (((val1 << j) & hashvalues[i].first) ? 1: -1) * hashvalues[i].second;
        }
    }
    retval = 0;
    for(size_t j = 0; j < LEN; j++)
    {
        if(weights[j] > 0.0)
        { retval |= (val1 << j); }
    }
    return true;
}

```

图 5.5 关键词签名实现的关键代码

Figure 5.5 The keyword signature implementation source codes

在图 5.5 代码中，getKeywordSimHash() 函数实现了基于算法 1 的关键词签名的计算，在这个函数中，首先利用 keywordHash() 函数对每个关键词进行了 Jenkins 哈希，然后对每个哈希结果进行-1 到 1 映射，最后将所有关键词的映射结果按位相加并和 0 比较，得到所有关键词的 SimHash 签名。

```

bool extractAndHash(const string& text, size_t topN, vector<pair<uint64_t, double> >& res)
{
    vector<pair<string, double> > wordweights;
    if(!jieba_extractor.Extract(text, wordweights, topN))
    { return false; }
    res.resize(wordweights.size());
    for(size_t i = 0; i < res.size(); i++)
    {
        res[i].first= jenkins_hasher(wordweights[i].first.c_str(), wordweights[i].first.size(), 0);
        res[i].second = wordweights[i].second;
    }
    return true;
}

bool getStringSimHash(const string& text, size_t topN, uint64_t& retval)
{
    vector<pair<uint64_t, double> > hashvalues;
    if(!extractAndHash(text, topN, hashvalues))
    { return false; }
    vector<double> weights(LEN, 0.0);
    const uint64_t val1(1);
    for(size_t i = 0; i < hashvalues.size(); i++)
    {
        for(size_t j = 0; j < LEN; j++)
        {
            weights[j] += (((val1 << j) & hashvalues[i].first) ? 1: -1) * hashvalues[i].second;
        }
    }
    retval = 0;
    for(size_t j = 0; j < LEN; j++)
    {
        if(weights[j] > 0.0)
        { retval |= (val1 << j); }
    }
    return true;
}

```

图 5.6 问句签名实现的关键代码

Figure 5.6 The question signature implementation source codes

在图 5.6 代码中，getStringSimHash() 函数实现了基于算法 2 的问句签名的计算，在这个函数中，首先利用 extractAndHash() 函数对问句进行关键词、权重提取以及每个关键词的 Jenkins 哈希计算，然后对每个关键词的哈希结果进行-1 到 1 映射并乘以关键词的权重，最后将所有关键词的映射结果按位相加并和 0 比较，得到所有关键词的 SimHash 签名。

点击问题后面的“查看详情”链接可以查看问题和答案的详细情况，如图 5.7 所示。在查看问题详情页面，可以点击“返回”链接返回问题管理模块主页面，教师和学生登录后问题描述后面的“回答”链接会显示为黄色，表示可以回答该问题，如果是教师或学生自己发布的问题，“编辑”问题的链接也会启用。目前管理员没有回答和编辑问题的功能权限。另外，如果学生的回答没有经过老师确认，在是否正确后面会显示“正确？”和“错误？”按钮，教师用户登录可以点击相应按钮进行答案确认，在确认后学生可以得到相应的积分。图 5.8 是教师用户点击“回答”链接后的问题回答页面，图 5.9 是问题编辑页面，目前回答和编辑问题的形式仅支持中文文字输入。学生用户提问页面和问题编辑页面实现结果类似，输入问题和科目提交即可完成提问，学生提问会写到 tab_waiting 表中，由管理员审核通过后转到 tab_question 表。教师提问会直接写到 tab_question 表。

问题详情						返回
提问者	王五	提问日期	2019/8/10			
科目	数学	浏览次数	5			
问题状态	已回答	被回答次数	2			
问题描述	二次函数的图像是什么?					回答 编辑
回答人 1	张三	回答日期	2019/8/11	是否正确	否	
回答信息	是一条直线					
回答人 2	叶老师	回答日期	2019/8/11	是否正确	是	
回答信息	是抛物线，有对称轴和顶点					

图 5.7 查看问题详情页面

Figure 5.7 Illustration of the problem details page

欢迎来到在线答疑系统

问题管理 常见问题管理 积分管理

欢迎您, 叶老师! [提问](#)

当前在“问题管理”模块 科目: 关键词检索: 问句检索:

问题回答 [返回](#)

提问者	王五	提问日期	2019/8/10
科目	数学	浏览次数	5
问题状态	未回答	被回答次数	1
问题描述	二次函数的图像是什么?		

请输入答案:

图 5.8 问题回答页面

Figure 5.8 Illustration of the answering page

欢迎来到在线答疑系统

问题管理 常见问题管理 积分管理

欢迎您, 叶老师! [提问](#)

当前在“问题管理”模块 科目: 关键词检索: 问句检索:

问题编辑 [返回](#)

问题科目	数学
问题描述	函数和映射之间的联系和区别是什么?

图 5.9 问题编辑页面

Figure 5.9 Illustration of the question editing page

常用问题管理功能和问题管理页面的对应功能实现结果类似, 在此就不一一展示。

5.1.4 积分管理模块实现

教师和学生用户登录后可以看到学生积分排名情况，学生积分排名功能实现结果如图 5.10 所示。管理员登录后在积分管理模块可以看到学生积分排名，也可以应学生要求对学生积分进行兑换，只需要点击学生信息后面的“兑换”链接即可，管理员也可以根据学号搜索对应的学生信息，然后进行积分兑换，积分兑换页面如图 5.11 所示，仅仅实现了积分扣减过程，奖品的选择则放在了线下由管理员控制。



排名	学号	姓名	班级	年级	学院	积分	操作
1	017020131	王五	1 班	2017	数学学院	20	兑换
2	017020105	小明	1 班	2017	数学学院	10	兑换
3	017020213	赵四	2 班	2017	数学学院	10	兑换
4	017050108	张三	1 班	2017	外国语学院	1	兑换
5	017020323	小虎	3 班	2017	数学学院	1	兑换

图 5.10 学生积分排名页面

Figure 5.10 Illustration of the students' points ranking page



积分兑换			
学号	017020131	姓名	王五
班级	1 班	年级	2017
学院	数学学院	积分	20
兑换积分			

提 交

图 5.11 积分兑换页面

Figure 5.11 Illustration of the point exchange page

5.2 系统测试

5.2.1 测试目的

软件测试是软件开发周期中的重要组成部分，其目的在于尽早发现开发过程中存在的错误或不足，本文从用户角度对系统的各项功能进行了测试，以便于开发得到交互友好、质量过关的实用系统。本系统部署在联想 ThinkSystem ST558 服务器上，服务器内存容量 32GB，2 颗 6 核 CPU，硬盘容量 2.4TB，服务器操作系统为 Microsoft Server 2016，系统后台数据库选用 MySQL 8.0.1，WEB 服务管理器为 IIS。

5.2.2 系统功能测试

系统功能测试主要尝试查找出几种类型的错误：界面出错、初始化及终止错误、功能缺失、性能错误、数据结构及数据库访问错误等。表 5.1 到表 5.4 是待审核问题模块、教师管理模块、问题管理模块和积分兑换模块的测试结果，其他相似模块的功能测试结果均通过。

表 5.1 待审核问题模块测试表

Table 5.1 Test table of the questions-to-be-audited module

功能项	测试内容	期望结果	测试结果
待审核问题显示	学生发布的待审核问题能否显示在该页面，并且问题相关信息能否正确显示	能够正确显示各项信息	通过
审核功能	审核通过按钮和不通过按钮能否正常使用	按钮功能正常	通过
审核后的问题发布	审核通过后，相关问题能够正确显示在问题管理页面，审核不通过则不显示。审核后问题不再显示在待审核页面	显示结果正常	通过

表 5.2 教师管理模块测试表

Table 5.2 Test table of the teacher management module

功能项	测试内容	期望结果	测试结果
教师相关信息显示	能够正确显示所有教师的相关信息	能够正确显示各项信息	通过
检索功能	能够根据工号进行相关教师搜索，并且搜索结果显示正常	能够搜索且显示正确	通过
教师信息编辑功能	管理员能够编辑教师的相关信息，能够删除教师信息	能够正常操作	通过

表 5.3 问题管理模块测试表

Table 5.3 Test table of the problem management module

功能项	测试内容	期望结果	测试结果
问题发布	教师和学生用户能够正确发布问题，发布后能够显示在正确的模块下	能够正确显示各项信息	通过
问题编辑	教师和学生可以正常编辑自己所提的问题	能够正常工作	通过
问题回答	教师和学生可以回答相关问题，回答之后相关信息能够正确显示	能够正确回答和显示	通过
问题检索	根据科目、关键字和句子能够较好检索到相关问题	能够检索到相关问题	通过
问题浏览	问题浏览功能能够正常显示相关信息	能够正常显示	通过
查看详情	查看详情链接能够正常使用，查看详情页面能够正确显示	能够正常使用和显示	通过
详情页面链接功能	查看详情页面的回答、编辑、确认答案等链接能够正常操作	能够正常操作	通过

表 5.4 积分管理模块测试表

Table 5.4 Test table of the point management module

功能项	测试内容	期望结果	测试结果
学生积分显示	能够正确显示学生信息和积分信息	能够正确显示各项信息	通过
积分兑换	积分兑换链接能够正常使用，能够正确进行积分兑换	能够正常使用	通过
积分兑换后排名	积分兑换后能够按照现有积分和兑换积分的和排名显示总积分	能够正常显示	通过

以上对系统详细的测试与分析结果表明，无论从功能上、自动检索精度上，都能满足系统的基本功能要求，系统使用操作上也符合用户正常的操作习惯，达到了上线运行的基本要求。

5.2.3 系统运行及维护

针对本系统制定的运行与维护的方法和手段如下：

1. 保管软件文档

软件文档是软件后续修改和维护的重要关键依据，丢失软件文档的后果是不堪设想的。因此在本系统的开发和试运行过程中，保存了软件需求规范、系统纲要设计规范和详细设计规范、源文件等，并且当软件在试运行过程中需要对文档或系统进行修改时，也对这些文档做了相应的补充说明，确保了软件和文档的一致性。

2. 保留开发环境

为了确保对软件的修改或补充能在当时的开发环境中进行，保留开发环境就非常必要。所以在系统运行和维护之前，也对系统开发环境进行了保存。

3. 建立维护手册

建立维护手册是软件维护中的必要过程。这就包括问题发现的时间和相应症状的描述、修改过程的描述和位置以及对修改问题的编号、对修改的模块名称的标识。本系统在试运行过程中，严格按照软件运行和维护过程要求进行记录，目前还在试运行阶段，后期如果修改补充的次数达到一定数量时，会考虑升级软件版本号，以体现对用户反馈意见的重视和软件不断完善的成果。

5.3 本章小结

本章详细介绍了系统实现结果，然后介绍了各个模块的功能测试结果，最后介绍了运行维护方面的注意事项。

第六章 总结与展望

6.1 总结

尽管可以利用 E-mail 和电子留言板的形式进行答疑解惑, 但解答时效完全依赖于答疑教师, 如果教师的解答不及时, 可能会影响学生的学习进展, 让学生失去学习的兴趣, 无法提升教学效果。本文描述的在线答疑系统正是在此背景下设计和实现的。本系统鼓励学生提问和参与回答问题, 能够增强师生之间、同学之间的互动, 增强学生学习的兴趣, 能够让学生及时解决自己的问题, 大大提升教学效果。

本文主要做了以下工作:

1. 根据中职师生的特点以及系统功能的需要进行了可行性分析, 确定了开发语言和数据库。
2. 为了减轻教师答疑压力, 本文研究了自动问答相关的理论与技术基础, 并实现了基于 SimHash 签名的相似问题检索方法。
3. 为了调动学生答疑解惑的积极性, 系统设计了积分功能, 提问和帮助回答问题都会得到相应积分, 积分排名和积分兑换功能能够激发学生的学习积极性, 进一步地促进了答疑系统作用的发挥。
4. 开发并实现了针对中职师生的在线答疑系统, 并进行了相关测试, 系统在运行性能和效果上都取得了较好的表现。

6.2 展望

当前实现的系统仍存在诸多不足之处, 例如无法通过语音、图像、视频等形式进行提问和回答; 没有课堂辅助功能, 例如上传课件、作业管理等; 内容审核功能也还没有引入自动内容过滤方法; 校外人员也无法自由注册使用本系统等。另外, 数据库的规范设计方面还需进一步改进。在未来的工作中会不断升级该系统, 不断完善其功能, 使得系统能够更好地为师生服务, 提升学校的教学质量。

参考文献

- [1]Simmons R F, Klein S, McConlogue K. Indexing and Dependency Logic for Answering English Questions[J]. American Documentation, 1964, 15(3):196-204.
- [2]Woods W A. The Lunar Sciences Natural Language Information System[J]. BBN report, 1972.
- [3]Kupiec J. MURAX: A Robust Linguistic Approach for Question Answering Using an On-line Encyclopedia[C], Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. 1993:181-190.
- [4]丁兴富. 远程教育学[M]. 北京: 北京师范大学出版社, 2001. 11
- [5]方利伟. 个性化在线答疑系统的研究与实现[D]. 浙江师范大学, 2007.
- [6]张银. 网络答疑系统的设计新思路及策略实现[J]. 现代教育技术, 2012, 22(5):88-92.
- [7]蒋昌金, 彭宏, 陈建超等. 基于组合词和同义词集的关键词提取算法[J]. 计算机应用研究, 2010, 27(8): 2853-2856.
- [8]朱甜甜. 短文本语义相似度量的方法和应用研究[D]. 华东师范大学, 2014.
- [9]伍浩铖. 社区问答搜索中排序方法的研究[D]. 中国科学技术大学, 2017.
- [10]陈伟鹤, 刘云. 基于词或词组长度和频数的短中文文本关键词提取算法[J]. 计算机科学, 2016, 43(12): 50-57.
- [11]张帅帅. 基于阅读理解的知识问答系统设计与实现[D]. 电子科技大学, 2018.
- [12]曹艳蓉. 基于中文社区的智能问答系统的设计与研究[D]. 南京邮电大学, 2018.
- [13]张苏可. 悟空问答社区系统服务端的设计与实现[D]. 南京大学, 2019.
- [14]桑志杰. 生成式问答系统技术研究与实现[D]. 北京邮电大学, 2019.
- [15]申豪杰. 基于知识图谱的电影知识问答系统研究与实现[D]. 重庆师范大学, 2019.
- [16]刘珮. 面向阅读理解任务的问答系统研究与实现[D]. 北京邮电大学, 2019.
- [17]<https://www.apple.com/siri/>
- [18]<https://www.microsoft.com/en-us/cortana>

- [19]<http://www.msxiaobing.com/>
- [20]https://dumall.baidu.com/?utm_source=xdzj&utm_medium=officialweb
- [21]Bordes A, Chopra S, Weston J. Question Answering with Subgraph Embedding[C], Proceeding of the 2014 Conference on Empirical Methods in Natural Language Processing. 2014:615-620.
- [22]Mikolov T, Sutskever I, Chen K, et al. Distributed Representations of Words and Phrases and their Compositionality[J]. Advances in Neural Information Processing Systems, 2013:3111-3119.
- [23]Yih W, He X, Meek C. Semantic Parsing for Single-Relation Question Answering[C], Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: Vol 2. 2014:643-648.
- [24]Dong L, Wei F, Zhou M, et al. Question Answering over Freebase with Multi-Column Convolutional Neural Networks[C], Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing: Vol 1. 2015:260-269.
- [25]徐晓芳. 基于条件随机场的中文分词技术的研究与实现[D]. 南京邮电大学, 2018.
- [26]黄丹丹. 基于深度学习的中文分词和关键词抽取模型研究[D]. 北京邮电大学, 2019.
- [27]程志远. 基于神经网络的中文分词研究[D]. 郑州大学, 2019.
- [28]<https://github.com/yanyiwu/cppjieba>
- [29]Thushara M G, Krishnapriya M S, Nair S S. A model for auto-tagging of research papers based on keyphrase extraction methods[C]. 2017 international Conference on Advances in Computing, Communications and Informatics(ICACCI). 2017:1695-1700.
- [30]Mihalcea R, Tarau P. TextRank: Bringing order into text[C]//Proceedings of the 2004 conference on empirical methods in natural language processing. 2004: 404-411.
- [31]Sadowski C, Levin G. Simhash: Hash-based similarity detection[J]. Technical report, Google, 2007.
- [32]https://en.wikipedia.org/wiki/Jenkins_hash_function

致谢

感谢我的导师张献国老师，在他的悉心指导下我的论文终于顺利完成了。在我撰写论文的过程中，从选题方向到系统设计，再到系统实现和论文定稿，张老师都耐心细致的解答我提出的各种细节问题，为此，张老师付出了大量精力，也牺牲了不少休息时间。张老师这种对学生认真负责的态度和严谨的治学精神是我今后在教科研工作中遵循的榜样。张老师给予我的帮助，令我终生受益，在此对张老师表示诚挚的谢意和崇高的敬意。

特别感谢邵允学老师给予我的无私帮助和默默支持，每每想起邵老师对吾辈这样愚笨至此的学生的关爱，心中都会涌起无尽感激之情。感谢魏宏喜老师，在我选择开题方向感到迷茫无助的时候，给予我宝贵的开题意见。还要感谢陈海涛老师，论文得以顺利完成，离不开陈老师的热情帮助。

在论文即将完成之际，灯下静坐，屋内传来父母和妻儿沉静均匀的呼吸声。想到学习期间无数个这样夜深人静的时刻，白天工作的喧嚣和夜晚孤灯下的苦读构成了我整个求学的画面，心中深深感激家人对我求学的支持和付出。在我读书这段时间，正值我的妻子怀孕和小儿出生，这个期间为了尽量不影响到我的学业，妻子、父亲、母亲，乃至丈人一家，都给予了无比的付出和支持，心中感激，无以言尽。

谨以此致谢献给所有关心、支持、帮助过我的老师和同学们！