

厦门大学学位论文原创性声明

本人呈交的学位论文是本人在导师指导下，独立完成的研究成果。本人在论文写作中参考其他个人或集体已经发表的研究成果，均在文中以适当方式明确标明，并符合法律规范和《厦门大学研究生学术活动规范（试行）》。

另外，该学位论文为()课题(组)的研究成果，获得()课题(组)经费或实验室的资助，在()实验室完成。(请在以上括号内填写课题或课题组负责人或实验室名称，未有此项声明内容的，可以不作特别声明。)

声明人(签名): 林
2016 年 11 月 29 日



厦门大学学位论文著作权使用声明

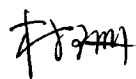
本人同意厦门大学根据《中华人民共和国学位条例暂行实施办法》等规定保留和使用此学位论文，并向主管部门或其指定机构送交学位论文（包括纸质版和电子版），允许学位论文进入厦门大学图书馆及其数据库被查阅、借阅。本人同意厦门大学将学位论文加入全国博士、硕士学位论文共建单位数据库进行检索，将学位论文的标题和摘要汇编出版，采用影印、缩印或者其它方式合理复制学位论文。

本学位论文属于：

（ ） 1. 经厦门大学保密委员会审查核定的保密学位论文，
于 年 月 日解密，解密后适用上述授权。

（ ☒ ） 2. 不保密，适用上述授权。

（请在以上相应括号内打√。或填上相应内容。保密学位论文应是已经厦门大学保密委员会审定过的学位论文，未经厦门大学保密委员会审定的学位论文均为公开学位论文。此声明栏不填写的，默认为公开学位论文，均适用上述授权。）

声明人（签名）： 

2016 年 11 月 29 日

摘要

互联网技术的发展推动远程教育模式应运而生,通过有机结合互联网技术和计算机辅助教学技术进行教学改革,使得远程教育模式渐成现代教育发展新趋势。智能答疑作为远程教育系统的重要部分,是用来评估远程教育系统优劣性的重要指标,其打破传统的时空限制,进行在线实时答疑。智能答疑综合运用自然语言处理、信息检索等多学科知识,为教师与学生间的交流互动提供了保证,对素质教育的实施及教学改革的推进有一定裨益。

本文首先介绍系统开发的背景、目的和意义,并提出智能答疑系统的解决方案。其次,详细分析系统的用户需求、功能需求和非功能需求,构建完善的体系结构,设计数据库表结构。接着,详细研究在智能答疑模块中如何有效应用中文信息处理技术,比如自动分词技术、句子语义相似度计算等。最后,为提高 FAQ 库的查询效率,本文进一步研究数据库快速定位技术,通过比较正排索引结构和倒排索引结构,结合智能答疑系统所涉及的数据特点,最终决定采用倒排索引结构来进行 FAQ 库的检索,提高关键词检索的效率和精确度。

本文设计并开发的智能答疑系统是搭建在微软公司开发的 ASP.NET 平台上,在开始开发系统功能模块前,先配置 ASP.NET 开发环境,为 WEB 项目的开发做准备。研究 MVC 设计模式和 ASP.NET MVC 框架,采用浏览器/服务器体系结构进行智能答疑系统各模块的设计和实现,并对智能答疑系统进行详细的测试,编写测试用例,保证系统达到需求和系统设计的要求。

关键词: 中文分词; 智能答疑; 远程教育

Abstract

With the development of Internet technology, distance-education mode, which effectively combines Internet technology and computer aided instruction to reform teaching mode, comes into being. It has gradually promoted the distance-education mode to be a new trend of the development of modern education. As a key part in distance-education mode, intelligent question answering, which is an important factor used to evaluate the robustness of a distance education system, breaks time and space limitations which exist in traditional question answering way and it deals with online and real-time answering. Intelligent-answering involves various aspects of multi-disciplinary knowledge and comprehensively uses natural language processing and information retrieval, providing high-quality interaction between teachers and students and playing a positive role in promoting the implementation of quality education and teaching reform.

First of all, we clarify the research background, purpose and significance and propose the solution of intelligent-answering system. Second, we analyze user requirements, functional requirements and non-functional requirements in detail, building complete architecture and designing the database structure further. Third, we study how to apply the Chinese information processing in intelligent-answering system, like Chinese word segmentation and sentence semantic similarity computation.

In addition, to effectively improve efficiency of query FAQ library, this dissertation further researches quickly positioning technology. Contrast index and inverted index and a detailed analysis of data characteristics, this dissertation finally decides to take the inverted index structure to improve accuracy and efficiency of keyword search.

In this dissertation, the IQAS is built on the platform .NET, based on MVC design pattern and adopting B/S architecture for the development, implement and detailed tests of the functional modules in the system.

Key Words: Chinese Word Segmentation; Intelligent Question Answering; Distance Education

目录

第 1 章 引言	1
1.1 研究背景	1
1.2 研究目的和意义	3
1.2.1 研究目的	3
1.2.2 研究意义	3
1.3 远程教育系统研究概述	3
1.3.1 发展历程和现状	3
1.3.2 系统的特点	4
1.4 智能答疑系统研究概述	6
1.4.1 概述	6
1.4.2 研究现状	6
1.5 本文主要研究内容	8
1.6 论文组织结构	9
第 2 章 开发平台及相关技术介绍	10
2.1 开发平台	10
2.1.1 Visual Studio 2010	10
2.1.2 Sql Server 2008	10
2.1.3 IIS 7.0	11
2.2 相关技术	11
2.2.1 B/S 体系结构	11
2.2.2 C#	12
2.2.3 ASP.NET	12
2.2.4 MVC 设计模式	13
2.2.5 ASP.NET MVC	15
2.2.6 LINQ To SQL 技术	16
2.3 本章小结	17
第 3 章 需求分析	18

3.1 系统目标	18
3.2 用户需求分析	18
3.3 系统功能需求分析	20
3.4 系统非功能需求分析	22
3.5 系统关键技术需求分析	22
3.6 本章小结	23
第 4 章 系统设计	24
4.1 系统体系结构	24
4.2 登录模块设计	26
4.3 智能答疑模块设计	27
4.3.1 快速中文分词机制	27
4.3.2 基于向量空间模型的句子相似度计算	30
4.3.3 基于《同义词词林》的语义相似度计算	32
4.3.4 索引结构	34
4.4 数据库设计	37
4.4.1 逻辑关系图 (E-R 图)	37
4.4.2 数据库表结构设计	37
4.5 本章小结	39
第 5 章 系统实现与测试	40
5.1 系统三层体系结构的实现	40
5.1.1 Model 部分的实现	40
5.1.2 View 部分的实现	41
5.1.3 Controller 部分的实现	42
5.2 登录模块实现	44
5.3 智能答疑模块实现	46
5.4 系统测试	53
5.4.1 中文分词测试	53
5.4.2 中文句子相似度测试	54
5.5 本章小结	55

第 6 章 总结与展望	56
6.1 总结	56
6.2 展望	56
参考文献	57
致谢	59

Contents

Chapter 1 Introduction	1
1.1 Research Background.....	1
1.2 Research Purpose and Significance.....	3
1.2.1 Research Purpose	3
1.2.2 Research Significance	3
1.3 Research Overview of Tele-Education System	3
1.3.1 Development and Situation.....	3
1.3.2 Characteristics of System.....	4
1.4 Research Overview of Intelligent Question Answering System.....	6
1.4.1 Research Overview	6
1.4.2 Research Situation	6
1.5 Main Research Contents	8
1.6 Dissertation Structure.....	9
Chapter 2 Development Platform and Related Technologies	10
2.1 Development Platform.....	10
2.1.1 Visual Studio 2010	10
2.1.2 Sql Server 2008.....	10
2.1.3 IIS 7.0.....	11
2.2 Related Technologies.....	11
2.2.1 B/S Architecture.....	11
2.2.2 C#.....	12
2.2.3 ASP.NET	12
2.2.4 MVC Design Pattern.....	13
2.2.5 ASP.NET MVC	15
2.2.6 LINQ To SQL	16
2.3 Summary.....	17
Chapter 3 Requirements Analysis	18

3.1 System Objective.....	18
3.2 User Requirements Analysis	18
3.3 Functional Requirements Analysis.....	20
3.4 Non-Functional Requirements Analysis	22
3.5 Key Technologies Requirements Analysis.....	22
3.6 Summary.....	23
Chapter 4 Systematic Design	24
4.1 Architecture Structure.....	24
4.2 Login Module Design.....	26
4.3 Intelligent Question Answering Module Design.....	27
4.3.1 Fast Chinese Word Segmentation	27
4.3.2 Sentence Similarity Computation Based on Vector Space Model.....	30
4.3.3 Semantic Similarity Computation Based on Chinese Thesaurus-Tongyici Cilin.....	32
4.3.4 Index Structure.....	34
4.4 Database Design	37
4.4.1 Entity-Relation Diagrams	37
4.4.2 Database Tables' Structure Design	37
4.5 Summary.....	39
Chapter 5 System Implementation and Test	40
5.1 Implementation of Three-Layer Architecture.....	40
5.1.1 Implementation of Model	40
5.1.2 Implementation of View.....	41
5.1.3 Implementation of Controller	42
5.2 Implementation of Login Module.....	44
5.3 Implementation of Intelligent Question Answering Module.....	46
5.4 System Test	53
5.4.1 Chinese Word Segmentation Test	53
5.4.2 Chinese Sentence Similarity Test.....	54

5.5 Summary.....	55
Chapter 6 Conclusion and Prospect	56
6.1 Conclusion	56
6.2 Prospect.....	56
References	57
Acknowledgement	59

第1章 引言

1.1 研究背景

互联网诞生大大推动了软硬件技术、网络技术等计算机技术的快速发展,尤其是“互联网+”概念的提出,使得互联网在教育领域方面的应用愈加受到关注,远程教育应运而生,逐渐成为提高国民素质、普及文化素养的主阵地^[1]。远程教育模式打破传统束缚,摆脱教育领域在时间、空间上的局限,为教师和学生提供了双重便利。远程教育系统中一个重要的模块就是在线智能答疑模块,其设计的优劣是评估一个远程教育系统好坏以及等次最重要的指标^[2]。在线智能答疑模块综合应用了多个学科的专业知识,包括中文分词技术、信息检索技术、句子相似度分析技术、语义相似度分析技术等,通过这些技术对问题进行在线分析,并自动检索问题库给出匹配度最高的答案。其效率极高、操作极简,大大提高了远程教育系统的质量,提高了师生间的互动效率,保证远程教育能够高效有序地运行,为教育发展和教学改革增添光彩。

虽然远程教育模式是一种全新的并具有广阔前景的教学模式,但其诞生发展至今并不是一帆风顺的,其依然存在很多关键性的难题亟待解决。首先是时间差异性,师生都有很多的私人时间,这些私人时间在绝大多数情况下都是不交叉的,学生不可能占用教师私人时间来提问,教师也不可能占用学生私人时间来解答,这样导致答疑的时效滞后性,学生只好花费更多的时间另辟蹊径去解决难题;其次是空间差异性,由于地理空间上的局限,有时师生间可能相隔数千里,教师无法第一时间为学生提供面对面的答疑解惑。然而,研究表明,当学生面对难题,能在第一时间获得解答是非常有必要的,这样才能增强学生对所遇难题的印象,所以在远程教育系统中在线智能答疑模块已然成为不可缺少的关键性模块,尤其是在这种时空冲突的情况下,更是必不可少。通过阅读相关中外大量文献发现,不管是国内还是国外,大部分现有的远程教育系统在设计在线智能答疑模块时都有一些考虑不是很到位、不是很周到的地方,总结归纳国内外现有的答疑模块,可以分为以下几种类型^[3-5]。

(1) 自主灵活、功能完善。这类答疑模块自带有齐全的词库、句库等跟智能答疑模块相关的信息数据库,学生在遇到疑问后向系统提交问题,智能答疑模块自动解析问题,提取出问题中的关键词、关键句,并与已有的、存储在系统端

的数据库进行相似度匹配,将所匹配到问题的答案返回给学生用户,并将问题及答案更新至相应的词库、句库中。当然,如果在系统数据库中匹配不到相似的问题和答案,系统会自动将学生提交的问题上传给专家,等待专家对学生问题作出相应的回答。这类智能答疑模块综合应用了计算机领域中多个相关子领域的技术,包括数据挖掘、中文分词、动态检索等。

(2) 弱智能化,功能单一。这类答疑模块没在本地构建强大的词库、句库等相关的信息数据库,学生在遇到疑问后,无法第一时间获得有帮助的解答,显现出系统高度的弱智能化,功能单一。这类答疑模块需要基于强大的人力支撑,系统管理员或教师需要 24 小时无间断处于在线状态,才能在第一时间为学生答疑解惑,对时间有强烈的束缚。这类答疑模块模式有点像网络贴吧论坛的讨论模式,局限性强,且目前市场上绝大多数的答疑模块都属于这类。

(3) 弱电子化,功能缺失。这类答疑模块模式比较呆板,师生之间交流互动的方式仅限于系统公告、留言板、短信等,具有严重的滞后性。学生在遇到疑问后,无法第一时间获得有帮助的解答,使得许多教学信息、教学情况以及教学效果不能得到及时反馈。当然,虽说这类答疑模块已逐渐退出市场,但不可否认,它还是存在于一些网站、论坛、贴吧中的。

从以上文献资料中归纳的现有答疑模块可以发现,不管是国内还是国外,大部分现有的在线答疑模块都属于以上其中一种,详细分析这些答疑模块发现,这些答疑模块存在一些共性的缺点亟待解决,主要表现为:

(1) 答疑模块智能化程度低。以上介绍的几种答疑模块智能化程度不一,有的可以一定程度地完成自动答疑,有的可以完成人工前提下的在线答疑,而有时在智能性和实时性方面都有缺失,总而言之,都不同程度地缺乏一定智能性。而且,智能答疑模块解析语句、关键词的方式相对比较单一,系统对用户的计算机操作基础要求高。同时,用户对系统的评价渠道不畅通,不能及时地对系统体验进行反馈,造成用户体验性较差。

(2) 答疑的方式比较单一。以上介绍的几种答疑模块在设计答疑方式时比较单一,不够丰富,要么是“学生提问-教师回答”的模式,要么就是“关键词匹配数据库”的模式进行答疑,答疑方式比较单一,无法给学生提供多样性的选择。然而,在远程教育模式下,一个优秀的智能答疑系统应该不仅仅能够让系统通过

自动检索来匹配问题和答案,并且能够在学生有需要的情况下,让教师对学生的问题作出进一步的解答和说明。

纵观国内和国外的智能答疑模块,绝大部分的在线答疑都不可否认地存在些许不是很到位的问题和不足,鉴于此,研究并设计了远程教育模式下智能答疑系统,来完善现有大部分答疑系统所存在的不足。

1.2 研究目的和意义

1.2.1 研究目的

通过阅读大量国内外相关文献,研究现有市场上所有答疑系统的大致分类,分析这些答疑系统所存在的一些共性的缺点和不足。有效结合了“学生提问-教师回答”模式和“关键词匹配数据库”模式,综合运用了数据挖掘、中文分词、动态检索、自然语言处理等计算机领域的相关技术^[6-7],实现了FAQ库自动更新、自动查找匹配问题答案、关键词提取、自然语言解析等功能,力求研究、设计并开发出一套可靠、安全、易用、简便的远程教育智能答疑系统,使得答疑系统不仅能够通过自动检索来匹配问题和答案,并且能够在学生有需要的情况下,让教师对学生的问题作出进一步的解答和说明。

1.2.2 研究意义

为了更好地推广践行“互联网+”理念,远程教育模式作为网络教育及应用分支中一个新兴的热点问题,打破传统束缚,摆脱教育领域在时间、空间上的局限,为教师和学生提供了双重便利。研究、设计并实现远程教育系统,其智能答疑模块的智能化程度早已成为不可不考虑的重要因子。而且,伴随着数据挖掘、中文分词、动态检索、自然语言处理等计算机相关技术的不断发展,远程教育模式终将成为教育改革的必然趋势,所以智能答疑系统具有非常广阔的应用前景。

1.3 远程教育系统研究概述

1.3.1 发展历程和现状

学术界对远程教育进行研究的学者很多,而且近几年出现井喷之势,不同学者对远程教育的理解都不太一样,对远程教育所下的定义也不太一样。加利福尼亚州远程教育计划,简称CDLP,它对远程教育的理解是这样的:“远程教育,简称DL,其主要负责连接教育资源与学生之间的联系。”

远程教育的发展历程与计算机技术的发展历程比较同步,它起源于欧洲部分

发达国家以及美国。当时这些发达国家中都存在大量无法到当地学校就读的人,获得教育学习的机会不多,这些国家就开放当时世界上最先进的邮政系统让他们可以异地学习,这样使得家住偏远地方、不能去学校上学以及身体残疾的人可以远程获得教育。但当时这样模式应用不是很广泛,只在少数几个发达国家中应用,而且教育成本相对较高^[8-9]。

到了 19 世纪 80 年代,伊利诺伊卫斯理大学开发了第一套真正意义上的远程教育系统,并将这套系统用于对本校本科毕业生、硕士毕业生进行学位授予,使得他们不需要实地出席学位授予大会即可获得各自的学士和硕士毕业学位。至此,这种远程教育系统引起美国甚至全世界的关注,直到 19 世纪 90 年代,肖托夸运动又快速推动了这种远程教育的在线网络教育形式。

进入 20 世纪后,远程教育的发展遇到了新的机遇,20 世纪 20 年代出现的广播技术以及 20 世纪 40 年代出现的电视技术都为远程教育更新了崭新的通信形式。从事教育领域的人员可以利用广播和电视技术远程播放教育音频节目和视频节目,打破传播的学习模式,为无数潜在学生人员提供了良好的学习机会。

此外,在 20 世纪初期,长途电话系统发展迅猛,大大增加了参加远程教育学习的学习群体,但是相比于广播和电视技术,长途电话系统相对来说没有发挥非常重要的作用。后来,一些发达国家开发或引进了各自的电话会议技术,利用电话会议技术,教师和学生之间可以实时进行语音交谈,不管他们处于何地,都能很好地保证通话信息没有传输延迟。

除了邮政系统、广播技术、电视技术、长途电话技术、电话会议技术外,随后远程教育在发展过程中仍更新了不同的通信技术,来逐步提高教师和学生之间进行交流沟通的能力,逐步降低交流沟通所需成本。尤其是 20 世纪末计算机的出现,人们开始通过电话线连接计算机,网络的世界让师生之间的沟通交流畅通无阻,教师和学生只需要坐在会议室内就可以通过计算机进行通信和学习。计算机技术、网络通信技术的出现,教师和学生可以实时动态交换视频、音频、图片、问题等电子数据,使得在远程教育通信技术应用上,邮政系统、广播技术、电视技术、长途电话技术、电话会议技术等被逐步减少使用,甚至被淘汰。

1.3.2 系统的特点

(1) 可靠且安全。对于远程教育系统来说,可靠性和安全性是极其重要的,

流转于远程教育系统中的电子数据的安全性需要得到充分的考虑。优秀的系统,在设计和开发时,需要懂得保护数据信息,应具有数据备份和恢复功能,所有的 Cookie 需进行多层加密、数据流转需多层校验、数据表单需多次进行传输前检查,使得数据不管处于输入、提交、传输、存储等各个阶段都应该是高度安全的。

(2) 分布式管理。远程教育系统支持用户不同时间在多个不同的地方进行学习,系统建立一个核心主站以及多个枝状的地区分站,不管是教师还是学生都可以通过主站或者分站进行注册、登录、开展教务教学、记录学时、建立档案等操作,分布与各地区的分区定时将数据汇总至核心主站,实现实时数据交换,这样很大程度上增加了网络带宽,均衡了网络负载。

(3) 实时监控机制。远程教育系统具有实时监控的功能,系统通过一定的设计或设置防止学生挂机获取学时、同时进行多门课程学习等,这些消极的学习方式都可通过实时监控来预防,可以监控的条件与参数可以通过管理员在后台进行动态的设置修改,对允许、禁止学员学习的资源进行自定义设置,这样能够很好地满足各种不一样的教育、学习和培训需求。

(4) 具有多角色权限管理机制。远程教育系统的权限设计非常严格,不是说任何用户都可以访问系统中的任何资源,不同用户对应不同的角色,不同的角色被赋予不同的权限集,系统采用多角色权限管理机制,实施分级管理主要包括了信息数据统计、考试成绩管理、成员管理、课程资源管理等功能。

(5) 全程统计与跟踪功能。远程教育系统可以对学员的任何情况进行实时的跟踪和统计,管理员和教师可以实时知道参与远程教育学习的任何一个学员的在线测试情况、学习投入时间、学习完成量、学习进度等信息,加之配备有相应的反馈功能,全自动智能化操作避免了统计过程中浪费时间、出现错误等情况。

(6) 易用性。不仅仅是远程教育系统,在“服务至上”理念的高度催动下,任何一个计算机系统在研究、设计和开发过程中都要考虑到易用性要求,需满足系统中不同用户的不同需求,明晰框架,使得系统操作起来方便直观,且界面指示性强,充分考虑不同用户的行为特点,不同用户都能轻易上手,确保教师、学生都能无阻碍且得心应手地完成在线学习交流。

(7) 实时性。实时性是远程教育系统首要也是最为重要的特点,无论是在线交流,还是文本、音频、视频、图片等数据信息交换,都要保证高实时性。可

以说,如果一个远程教育系统缺乏实时性,它也就丧失了其本身的价值。

(8) 功能全面完善,可扩展性强。“麻雀虽小,五脏俱全”,不管是规模多大的远程教育系统都必然具有完善的教育教学管理功能,并提供业务及开发接口,能有效地去适应业务需求方面的变化,支持二次开发再利用,从而进一步保障远程教育系统的平滑升级,满足业务需求和管理模式的变化。

1.4 智能答疑系统研究概述

1.4.1 概述

随着各类教学信息的扩展积累、现代教育模式的发展以及计算机网络技术的不断更新,教育改革中面临的一项难题亟待解决,那就是生师比越来越大,学生人数越来越多,导致教育资源略显匮乏,传统的面授教学在这场教育改革中匍匐前行。现如今仍有不在少数的教育网络所设计的答疑系统模式仍比较呆板,师生之间交流互动的方式仅限于系统公告、电子邮件、留言板、短信等,具有严重的滞后性。学生在遇到疑问后,无法第一时间获得有帮助的解答,使得许多教学信息、教学情况以及教学效果不能得到及时反馈,延误教与学的进程。

智能答疑系统(英文拼写为 Intelligent Question Answering System)^[10]是远程教育模式中加强教师与学生之间交流沟通的重要组成部分,不仅是教师教学过程之中的有益补充,更是学生进行课后巩固、课后复习的重要途径。智能答疑系统以“问题-答案”一对一映射的方式,将问题和答案存储在系统数据库中,当学生向系统提交问题时,智能答疑系统自动利用自然语言技术解析问题,比较所提交问题和数据库中问题的相似度,并根据“问题-答案”的一对一映射提取出相应的答案,并返回给学生。同时,智能答疑系统也是帮助教师改善教学模式、改进教学方法的有力工具,其自带的分析统计功能,可以帮助教师记录每个学生的知识薄弱点,从而定期地对学生做出系统性的辅导和总结。智能答疑系统综合应用了人工智能技术、计算机技术、数据挖掘技术、网络通信技术、信息检索技术、自然语言处理技术等,是一个极具适应力的智能系统,对用户所提供的问题不做格式限制,支持自然语言式的提问,系统会智能地检索数据库中的问题,并返回问题答案,具有友好的人机交互性能、知识学习能力强、逻辑推理能力强、数据分析统计能力强、知识记忆能力强等优点。

1.4.2 研究现状

国内外有不少的教育研究机构和平台从事该类型应用型系统的研究,从以下几个方面来对比下国内外答疑系统的现有研究状况^[1]。

智能答疑系统的功能:国内大部分现有的智能答疑系统一般都具有解答评价、返回答案、个性化定制、用户管理、搜索匹配问题、提出并提交问题等功能,并提供形式丰富多样的信息数据查询机制、数据深度挖掘技术、数据统计分析等功能。相比之下,国外的智能答疑系统一般没有集成如此多样化的功能,功能虽然比较单一,但在智能答疑模块方面做得独具特色,其不仅构建自身资源库,还在此基础上,充分有效地利用网络第三方资源,大大提升了答疑效果和质量,比如国外的 START 系统以及 Ask Jeeves for Kids 系统等就是典型的代表。

智能答疑系统的独立性:国内大部分现有的智能答疑系统都是以模块化的形式,将其嵌入至其他系统中,所以经常又被称为智能答疑模块。几乎没有或者说很少有以相对独立的姿态来构建系统,与其他网络教学平台的结合,大大降低了它的可移植性,使其严重受到束缚,失去独立性。相比之下,国外的智能答疑系统一般都不附属于任何一个系统,都是以独立架构或独立系统存在,这种做法提高了系统的可移植性,降低了系统的耦合度。

智能答疑系统的智能性:其综合应用了人工智能技术、计算机技术、数据挖掘技术、网络通信技术、信息检索技术、自然语言处理技术等,其中自然语言处理技术以及信息检索技术是核心,然而国内大部分的智能答疑系统所应用的这两种技术仍不太成熟,还属于前期摸索阶段,而且国内大部分的智能答疑系统在反馈机制设计上仍不完善,欠缺一定的智能性。相比之下,许多国外发达国家在技术领域都领先与国内,虽然在自然语言处理技术和反馈机制方面也存在一定的问题,但是许多发达国家在设计开发智能答疑系统时增加了额外的答案处理模块和问题解释模块,更凸显人性化设计。

智能答疑系统的资源库:国内大部分现有的智能答疑系统是按章或者按节的方式来设计资源库的,每次学生提交的问题和系统或教师给出的答案,都会被系统以章节的形式存放在资源库,当下次再有学生提交相同问题或相似问题时,系统也只会缩小搜索范围,在特定的章节中进行搜索查找,而这类智能答疑系统最大的缺陷就是资源库中的数据不够丰富,过于单薄。相比之下,国外大部分智能答疑系统一般都拥有强大丰富的资源库数据,处于基于自己构建资源库外,它们

还定期搜集网络第三方资源库,做到更加丰富、更加多样,避免了国内智能答疑系统资源库匮乏的问题。

智能答疑系统的技术瓶颈:可以说相比于国外智能答疑系统,国内的智能答疑系统在研发阶段所遇到最大且首要的技术瓶颈问题就是中文信息处理技术,国外许多国家以英语为母语,所以英文信息处理技术相对比较成熟,也有较为成熟的统一标准。而国内的智能答疑系统所要处理的都是中文信息,中文词库的匮乏、中文语言上的歧义都严重影响了语句进行中文分词时的准确度。此外,信息检索技术、数据挖掘技术、机器学习技术也是国外发达国家走在相对比较前列,国内技术的不成熟很大程度地影响了智能答疑系统的前进^[12]。

1.5 本文主要研究内容

通过大量阅读国内外相关文献,研究现有市场上所有答疑系统的大致分类,分析这些答疑系统所存在的一些共性的缺点和不足,有效结合了“学生提问-教师回答”模式和“关键词匹配数据库”模式,综合运用了数据挖掘、中文分词、动态检索、自然语言处理等计算机领域的相关技术,实现了FAQ库自动更新、自动查找匹配问题答案、关键词提取、自然语言解析等功能,力求研究、设计并开发出一套可靠、安全、易用、简便的远程教育智能答疑系统,使得答疑系统不仅仅能够通过自动检索来匹配问题和答案,并且能够在学生有需要的情况下,让教师对学生的问题作出进一步的解答和说明^[13]。具体的研究内容如下。

(1) 参考国内外现有的智能答疑系统的框架,提出远程教育模式下智能答疑系统的解决方案。阅读大量国内外的相关文献,理清智能答疑系统的设计开发思路,详细阐述系统开发的背景、目的和意义,并对远程教育模式和智能答疑系统的发展历程和发展现状加以介绍。

(2) 详尽分析系统需求及工作流程,设计系统数据库。一是要弄清使用智能答疑系统的主要用户有哪些,这些用户应该分别对应什么角色,每个角色应被赋予怎样的权限;二是要弄清要设计开发的智能答疑系统应具备哪些功能,以充分满足远程教育的需要,从而建立完善的系统架构体系;三是详尽分析智能答疑系统应达到怎样的特性,也就是系统的非功能需求。

(3) 研究在智能答疑系统中如何应用中文处理技术。设计开发的智能答疑系统所要处理的都是中文信息,中文词库的匮乏、中文语言上的歧义都严重影响

了语句进行中文分词时的准确度。研究在智能答疑系统中如何应用中文处理技术,研究如何进行中文信息处理、语句相似度的计算以及句子间语义相似度计算等。

(4) 研究数据库快速检索技术,提高数据访问、提取效率。本系统采用倒排索引结构来对系统数据库进行快速检索,力求提高查询关键词的检索精确度和效率。

(5) 搭建平台,配置环境。本文的智能答疑系统是搭建在.NET 平台上的,在开发系统前,需首先配置基于.NET 的 Web 服务器环境,确定进行系统开发所采用的 MVC 模式,并应用浏览器/服务器模式对智能答疑系统各个功能模块进行设计、开发和实现。最后并对智能答疑系统进行详细地性能测试、功能测试等,编写具有代表性的测试用例,从而确保设计开发的系统能够充分满足系统需求。

1.6 论文组织结构

第1章引言。本章介绍了本智能答疑系统的研究意义、研究意义和研究背景,对智能答疑系统和远程教育模式的相关概念和发展历程进行简要的介绍,介绍了本文主要研究内容及论文的框架。

第2章开发平台及相关技术介绍。先简要介绍了采用什么类型的服务器、开发平台以及后台数据库,然后对 ASP.NET 平台、MVC 设计模式、B/S 体系结构简单地进行阐述,有了整体的框架思路,为后续的系统设计开发做好准备。

第3章需求分析。本章主要对智能答疑系统的需求进行了分析,详细阐述了系统用户需求、功能需求、非功能需求以及开发目标等。

第4章系统设计。本章对智能答疑系统进行了总体设计,阐述了系统功能及其架构,对智能答疑系统的数据库表结构和重要的功能模块进行详细地设计。

第5章智能答疑系统的实现与测试。本章基于前几章中系统开发采用的设计思想和设计模式对系统的 MVC 框架进行搭建以及各功能模块的实现。并对系统进行了详细完整的测试。

第6章总结与展望。总结本文中介绍的智能答疑系统,并对系统中存在的问题和不足进行展望。

第2章 开发平台及相关技术介绍

本章主要介绍智能答疑系统在编码开发介绍所用到的主要框架、技术、平台，为后续系统的开发打下基础。

2.1 开发平台

2.1.1 Visual Studio 2010

Visual Studio 是微软公司为广大程序员提供的免费集成开发环境，通过其强大的语言服务可以允许多种不同的编程语言在其上编写运行，包括 F#、J#、C#.NET、VB.NET、C 以及 C++ 等，同时还支持 CSS、JavaScript、HTML、XHTML、XML 以及 XSLT 等脚本语言，这些编程语言可以大大简化 Web Service 和 Web 应用的开发，尽显 .NET 框架强大的功能。

Visual Studio 2010 继承了微软编程人员的大量心血，于 2010 年第二季度正式发布，据微软公司介绍，该版本相比于前一版本，大大降低了软件的复杂度和混乱度，且可移植性强，支持开发多系统的应用程序，同时支持包括 DB2、SQL Server、Oracle 等多种大型关系型数据库。

2.1.2 Sql Server 2008

SQL Server 数据库也是由微软公司开发的，是一种关系型数据库服务器系统，其负责为其他应用软件或应用程序检索、存储和提取数据。其强大之处更在于它是一款分布式数据库服务器，也就是说 SQL Server 可以部署在单台计算机上，同时也可以部署在多台计算机上，这多台计算机通过互联网进行数据交互，所采用的编写语言有 ANSI SQL、T-SQL 两种。

2008 年第三季度，微软公司终于在千呼万唤中退出 SQL Server 2008，对数据实现“三自”管理：自我维持、自我组织、自我调节。SQL Server 2008 可以存储各式各样的数据类型，包括视频、文本、音频、图片等多种格式，不管是结构化还是半结构化数据，在 SQL Server 2008 上都可实现无损存储。同时，它还支持对空间数据、文本、文件、时间、电子邮件、可扩展标记语言等多种数据的同步、共享、分析、搜索等操作。相比于以前几个版本，SQL Server 2008 具有更好的可扩展性，与其他第三方的开发包或程序能更好地进行集成。它还更新了索引算法和压缩算法，使得同等条件下检索相同数据时间更短，同等条件下存储相同数据所占用空间更小。

2.1.3 IIS 7.0

互联网信息服务，简称 IIS，是一款集 Gopher 服务器和 FTP 服务器于一身的万维网服务器，编程人员能够利用 ASP、Java、VBScript、JavaScript 等编程语言编写网络页面，并在 IIS 上进行发布。IIS 7.0 是 IIS 的最新版本，具有更好的扩展性，拥有扩展性良好的模块化架构，可在需要时动态添加特定功能，不需要时动态删除特定功能，十分灵活。

相比于 IIS 的前几个版本，IIS 7.0 拥有新的六大特性。

(1) IIS 7.0 拥有扩展性良好的模块化架构，可在需要时动态添加特定功能，不需要时动态删除特定功能，十分灵活。

(2) IIS 7.0 拥有一个统一标准的 HTTP 管道，能够很好地进行应用程序的本地管理，无论是基于 Web 的认证系统还是基于窗体的认证系统，都能在 IIS 7.0 上很好地运作。

(3) 在 IIS 7.0 上，任何用户都可以根据需要建立自己的 IHttpModule 以及 IHttpHandlers，并且将它们放入统一的管道中。

(4) IIS 7.0 充分吸取了 ASP.NET 技术在系统设置方面的优点，进行分布式的 XML 设置。

(5) IIS 7.0 在基于前几个旧的服务器版本，改善了在问题解答和问题诊断机制上存在的问题，包括了原先在问题解答和问题诊断机制上存在的跟踪功能漏洞和 Runtime 状态漏洞。

(6) IIS 7.0 采取的是面向任务的管理员用户阶段，直观、新颖且可扩展性强。

2.2 相关技术

2.2.1 B/S 体系结构

可以说，传统的系统体系结构设计几乎全部都是采用客户端/服务器（简称 C/S）模式，其他任何一种系统体系结构都不可能撼动它的王者地位。但是 C/S 模式最大的缺点就是所到之处必留痕迹，它要求所有需要采用 C/S 模式访问系统服务器的客户端都需安装上指定的客户端软件，大大增加了耦合性，尤其是在移动互联网兴盛的今天，C/S 模式已经不能很好地适应分布式办公、移动办公等渐趋流行的办公模式。

随着 Internet 技术、Web 技术的快速发展,这种 C/S 模式逐渐淡出市场,而出现一种新型的系统体系结构,叫做浏览器/服务器(简称 B/S)模式,其结构如图 2-1 所示^[14]。相比于 C/S 模式, B/S 模式实现了无需安装、无需卸载、无需维护,克服了传统 C/S 模式需安装上指定的客户端软件的缺点。B/S 模式成功运用多种脚本语言、借助浏览器技术,用简化的程序编码取代 C/S 模式下的指定客户端软件,简化了业务处理逻辑,具有负载小、选择多、成本低、扩展性强等优点,无可置疑地逐渐取缔 C/S 模式,成为当今系统体系结构设计的首选。

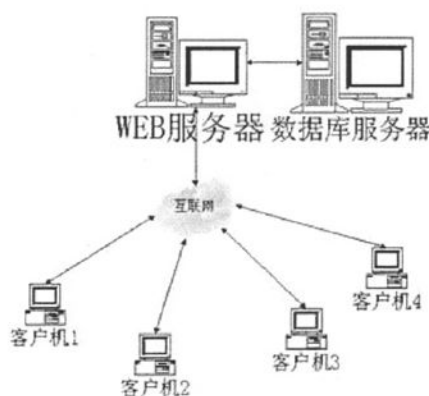


图 2-1 B/S 体系结构

2.2.2 C#

C#编程语言是一种多范型编程语言,是由 C、C++衍生而来,编码风格与 C、C++完全一样,但是相比于 C、C++, C#在集成其所有优点、改善其所有缺点的同时,集成了 C++语言在编译运行时高效率、VB 编程语言的可视化操作以及 Java 等面向对象编程语言的编程优点,所以一经推出很快就能被程序员熟悉,成为基于 .NET 平台进行系统开发的首选编程语言。C#编程语言有效地结合了快速应用开发技术的高生产性和 C、C++语言的原始优点,利用 C#进行编码开发时是本着面向组件、面向对象、泛型的、功能性的、声明性的、强类型的原则,具有面向对象、类型安全、通用、现代、简单等优点,专门用来进行软件系统的基础结构设计。

2.2.3 ASP.NET

动态服务器页面(简称 ASP)是在 IIS 2.0 版本上首次推出、在 IIS 3.0 版本

上发扬光大的。后来微软推出.NET 平台，ASP 发展成为 ASP.NET，虽然在语法上与 ASP 如出一辙，但相比于 ASP，ASP.NET 提供了全新的基础设施和编程模型，使用其开发的应用程序具有更好的稳定性和可扩展性。ASP.NET 是微软.NET 平台的一个重要部分，提供统一、定制化的 Web 开发模型，以方便构建企业级的 Web 应用程序。

ASP.NET 技术是基于微软公司研制开发的.NET 框架来建构 Web 应用程序的，其进行应用软件程序的开发可以放置在微软产品 Visual Studio 开发环境中，程序编码人员可以采用 Jscript、C#、VB 等多种与.NET 平台兼容的语言在 Visual Studio 上开发应用软件程序。

2.2.4 MVC 设计模式

模式-视图-控制器（简称 MVC）模式，是应用软件程序在设计开发过程中比较常用到的体系结构模式，尽管这种 MVC 体系结构模式由来已久，但一直受到“冷遇”，直到近几年 JSMVC、Monorail、CakePHP、Ruby on Rails 等 MVC 中几个核心框架的异军突起，才使 MVC 模式重新被拉回公众的视野，图 2-2 为体系结构图^[15]。从图 2-2 中可以看出，MVC 模式是将应用软件程序设计成三层模式，分别是表现层、数据层以及两者之间的交互层，这种隔离设计可以方便程序移植和扩展，并培养良好的编程习惯。简单地说，模型代表数据层，视图代表表现层，控制器代表两者之间的交互层，以便于两者之间的通信。在此详细介绍下三者之间的关系。

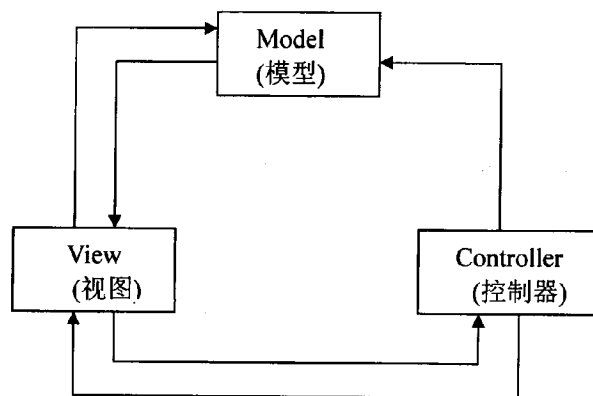


图 2-2 MVC 模式体系结构图

模型代表数据层,数据层是三层体系结构中的最底层,其中存储了流转于整个应用程序间的数据,这些数据包括业务数据库、用户信息、日志等,不管是文本数据、图片数据、音频数据还是视频数据等全部存储在数据层。除了数据以外,数据层还需负责与存储实际数据的数据库进行交互,并实现与数据库交互过程中所有的业务逻辑,包括增加、删除、修改、查询等操作,其通过建立关系-对象映射,将底层数据库表一一映射成相应的对象元素,从而简化了与数据库的交互操作,让应用程序开发人员无需编写复杂的 SQL 语句就能与数据库进行交互。在模型中编写的代码是为了实现业务逻辑功能,其包含了数据间交互的所有规则,所以可以说模型中业务逻辑的实现是整个 MVC 模式的核心部分。这样从表现层中剥离业务逻辑,大大降低了应用程序的耦合性,增强应用程序的可移植性和可重用性。

视图代表表现层,表现层是三层体系结构中的最上层,是模型的可视化表示,是直接为用户交互的一层,或者说是用户界面,用户应用程序所能看到的所有页面都属于表现层,其负责根据系统返回的数据以可视化的方式展现给用户看。在以前,所有的页面几乎都是静态的 HTML 文本页面,所有的页面不能进行动态修改,不能内嵌交互逻辑。而现在,表现层绝大多数情况下也是采用 HTML 页面,只是会内嵌部分简单的交互逻辑程序。当然我们必须承认,一个优秀的表现层设计应该尽可能少地包含业务逻辑交互,而将业务逻辑交互全部转交给模型和控制器,使视图专于其职,页面尽可能简单简化。视图在实现 HTML 页面自带功能之外,通常还会额外实现类、帮助器、格式化日期等其他功能。总而言之,视图即负责将数据以特定的可视化形式展现给用户。

控制器是三层体系结构的中间层,介于表现层和数据层之间,进行模型与视图之间的转换,起中间桥梁作用。控制器负责接收来自用户或视图的用户请求,明晰用户需要做什么、决定做什么,并将其转换成数据请求后发送给模型,与模型进行通信,最后并将模型返回的数据再次发送给用户或视图,可以说所有的业务逻辑处理都在控制器实现。控制器是由操作模型的方法和动作组成的,当系统用户打开应用程序的链接,链接自动触发请求,请求通过调度器进行转发。总而言之,控制器不涉及任何的用户界面、任何的数据操作以及任何的系统业务处理逻辑,它仅仅完成模型和视图之间的转换,并根据不同用户或视图的需求操

作返回不同的用户页面。

应用软件程序通过 MVC 模式可以实现更好地分层,将展示、数据及业务处理逻辑彻底隔离开,降低耦合度。利用 MVC 模式设计开发的应用软件程序具有以下优点:(1)有助于应用软件程序实行工程化管理;(2)使得应用软件程序中的控制层理念更加的明确;(3)增强了应用软件程序的可移植性;(4)提供程序接口,方便后期的扩展;(5)单个模型可以和多个视图进行匹配映射。

2.2.5 ASP.NET MVC

ASP.NET 框架一个十分重要的体系框架就是 ASP.NET MVC,它是 .NET 平台和 MVC 模式的有机结合,在 .NET 平台上以 MVC 模式进行 Web 应用软件程序的开发^[16]。它可以帮助开发人员构建更加便于维护的 Web 站点,以 MVC 模式降低 Web 站点在应用层间的依赖性。ASP.NET 技术大大提高了平台固有的支持测试驱动开发的测试性能,可进行完整的控制页面标记,这种对固有控件的剥离使得单个组件的应用可以更加灵活地进行使用、控制,使得用其开发的应用软件程序可以更加容易地进行测试、制定和修改。利用 ASP.NET MVC 技术构建的应用软件程序具有模块化的功能,这种模块化的功能让开发人员在开发过程中更具独立性,可以直接并行开发,而不必等待流水进行,大大提高了应用软件程序的开发效率。

ASP.NET MVC 体系结构的运行机制如图 2-3 所示,首先是由用户或视图向系统 Web 服务器或数据库服务器发送格式为“<http://HostName/ControllerName/ActionName/Parameters>”的链接请求,以待响应。该链接请求首先会被 ASP.NET 平台的路由映射机制所拦截,按照路由映射机制所指定的路由映射规则来解析链接请求中的 Controller、Action 以及相关的参数值。然后系统会根据解析出来的 ControllerName 在 Controllers 文件夹下查找名为 ControllerName 的控制器类,根据解析出来的 ActionName 在名为 ControllerName 的控制器类中查找名为 ActionName 的 Action 方法,并将传递过来的 Parameters 传递给 Action 方法中的相关参数,然后编译运行 Action 方法。在运行结束后,会将运行结果返回给相应的视图页面,并将控制器传递过来的 ViewData 数据一并传递给视图页面,其中包括了视图中页面显示所需数据以及控制视图显示所需的控制量。详细的运行机制请见图 2-3。



图 2-3 ASP.NET MVC 运行机制

ASP.NET MVC 体系结构的优点可以总结如下。

(1) 有效分离，耦合性低。对于 ASP.NET MVC 体系结构来说，整个体系结构的核心联系都体现在接口上，且都可以进行单独设计实现，所以使得每个功能都可以进行单独的单元测试，而不必一直反复地运行控制器。并且它支持 MS Test、MBUnit、NUnit 等任何一个单元测试框架进行系统测试，实现快速开发、快速测试。

(2) 可插拔性和可扩展性强。ASP.NET MVC 体系结构中所有的设计都是为了使得操作更加简便，简便地自定义、简便地更换，如参数序列、路由策略、视图引擎等。同时，它也支持进行控制反转、依赖注入等操作，比如 NHibernate、Spring Net、Windsor 技术等。

(3) ASP.NET MVC 体系结构包含功能强大的 URL 拦截映射机制，通过简单的 URL 编写规则简化复杂的程序流程化，使得程序编写起来、阅读起来更加简洁明了。而且，ASP.NET MVC 体系结构友好的命名模式使其很轻松地支持 REST、SEO 等扩展操作。

(4) 视图页面编写简便。ASP.NET MVC 体系结构支持丰富多样的视图模板，.ASP、.ASCX、.Master 等应有尽有，也就是说可以通过简单地本地化操作、数据绑定操作、模板、声明服务器控件、<%=%>代码嵌套片段、模板页等来快速简便地实现 ASP.NET 功能。

(5) ASP.NET MVC 体系结构完全延续并完全支持了 ASP.NET 的优秀技术，如程序体系化、系统配置、健康监测、profile/session 状态管理、数据缓存、角色/成员、URL 授权、Windows 身份验证、表单验证等功能。

2.2.6 LINQ To SQL 技术

语言集成查询 (Language Integrated Query, 简称 LINQ) 是 ASP.NET 体系框架下一种全新的数据查询方法, 其编写语法类似于 SQL 语句, 但是在 SQL 语句的基础上增加了对 ASP.NET 的特有支持, 通过 LINQ 编写的程序代码可以通过 .NET 平台编译运行后直接操作数据库。LINQ 技术提供了一大堆独特的查询操作符号, 可以用于对数据库、XML 文件等对象集合的查询, 通过 LINQ 处理引擎将编写的 LINQ 语句转换为可执行的语言代码, 进而对数据库进行操作。

LINQ To SQL 是 LINQ 技术的扩展, 它是 .NET Framework 3.5 发布的新组件, 在 LINQ 技术的基础上, 对 SQL 语句操作做了更多的功能扩展。简单地说, 它是一种“对象-关系”的映射模型, 可以帮助程序开发人员将关系型数据库对象转化为 .NET 对象, 并可以通过简单的方式管理关系型数据库对象与 .NET 对象之间的映射关系。通过映射后, 程序开发人员即可利用 LINQ To SQL 技术直接对关系型数据进行增加、删除、修改、查询等数据库操作, 通过当开发人员对 .NET 对象进行操作时, 系统也会自动实时将相应操作、相关数据更新至相应的数据库。

2.3 本章小结

本章介绍了智能答疑系统在编码开发介绍所用到的主要框架、技术、平台, 为后续系统的开发打下基础。

第3章 需求分析

智能答疑系统的开发要经历分析、设计、编码、测试四个阶段，这四个阶段对智能答疑系统来说都是非常重要的，上一章介绍了远程教育模式下智能答疑系统在开发过程中所需要用到技术和平台，本章将详细介绍智能答疑系统的需求，需求分析作为整个系统开发周期中的开端，其重要性不亚于其他任何一个阶段，是项目开发的基石。

3.1 系统目标

通过分析国内外现有智能答疑系统的缺点和不足，有效结合“学生提问-教师回答”模式和“关键词匹配数据库”模式，综合运用了数据挖掘、中文分词、动态检索、自然语言处理等计算机领域的相关技术，实现了FAQ库自动更新、自动查找匹配问题答案、关键词提取、自然语言解析等功能，力求研究、设计并开发出一套可靠、安全、易用、简便的远程教育智能答疑系统，使得答疑系统不仅仅能够通过自动检索来匹配问题和答案，并且能够在学生有需要的情况下，让教师对学生的问题作出进一步的解答和说明。从而打破传统束缚，摆脱教育领域在时间、空间上的局限，为教师和学生提供了双重便利，更好地推广践行“互联网+”理念。

3.2 用户需求分析

通过详细对比国内外的智能答疑系统，分析系统的实际需求，将本文中的智能答疑系统的系统用户分成三种类型，分别是教师用户、学生用户和游客用户。下面对三种不同类型的系统用户需求进行详细地分析。

(1) 游客用户。所谓游客，也就是指没有系统账号，更不能登录系统的用户，它是系统中最低层级的用户，拥有最小的权限集。游客用户在访问智能答疑系统时只能简单地通过客户端浏览器浏览系统常见问题、浏览问答列表，以及搜索一些自己想要查找的信息，除此之外，游客用户不再具有其他权限。图3-1为智能答疑系统中的游客用户用例图。

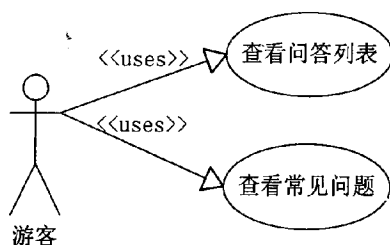


图 3-1 游客用户用例图

(2) 学生用户。学生用户是智能答疑系统所要服务的主要对象，是主要的一类系统用户，系统问题库的构建也全依赖学生源源不断提出的问题。所以，学生用户除了拥有游客用户所拥有的所有权限外，学生用户还可以向系统或教师提出问题、随时查看自己所提出问题的列表、对系统的智能回答或教师的回答作出反馈评价。另外，学生用户对系统界面的友好性、用户体验是否良好等要求较高，系统为学生用户提供资源是否全面、及时、准确、丰富也成为学生用户进行评价的重要因素。图 3-2 为智能答疑系统中的学生用户用例图。

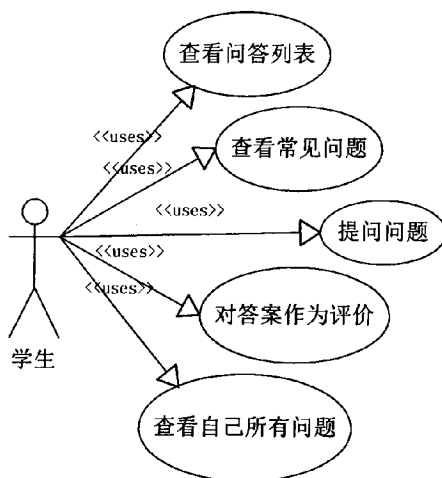


图 3-2 学生用户用例图

(3) 教师用户。教师用户也是系统的一类主要服务对象，为学生用户解答问题，是重要的系统用户，系统答案库的构建也全依赖教师用户为学生用户源源不断地进行解答。而且，答案库的初始构建也全依赖于教师用户提供数据，以及

对学生用户所提出问题、不满意或者低评价的回答重新解答，扩充并调整答案库资源。所以，教师用户除了拥有游客用户所有的权限、关心学生用户所在乎的用户体验度外，教师用户还关注内容是否完整是否准确、资源是否丰富以及问答库体系是否完善合理等。图 3-3 为智能答疑系统中的教师用户用例图。

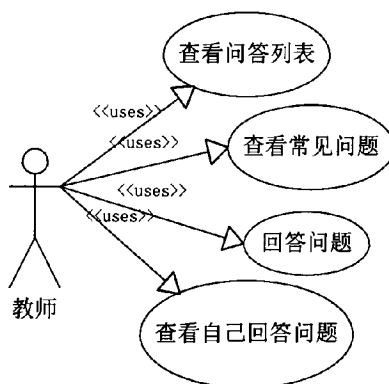


图 3-3 教师用户用例图

3.3 系统功能需求分析

从以上用户需求可以看出，远程教育模式下智能答疑系统应至少需实现以下功能，方能满足用户的需求。

(1) 查看问答列表。不管是游客用户、学生用户还是教师用户，都拥有查看问答列表的权限，用户访问系统后，系统会自动提取数据中的问答列表，以供他们查阅。除此之外，学生用户可以筛选出自己曾提出问题的列表，教师用户可以筛选出自己曾解答问题的列表。这些都是系统最基本的功能。

(2) 查看常见问题。所谓的常见问题，就是多名学生都曾提出该问题，由于智能答疑系统在智能性上的要求，智能答疑系统必须构建常见问题库（Frequently Asked Questions，简称 FAQ）。系统需能够自动统计 FAQ 中各类问题出现的频率，从高频至低频进行筛选。不管是游客用户、学生用户还是教师用户，都拥有查看常见问题的权限，据统计，至少一半以上的用户只需查看常见问题即可解除疑惑，大大减少了提问次数，提高了效率。

(3) 学生提出问题。学生用户是智能答疑系统所要服务的主要对象，是系

统另外一类主要的用户，系统问题库的构建也全依赖学生源源不断提出的问题。向系统或教师用户提出问题是学生用户最基本的功能需求，智能答疑系统应开放相关接口让学生用户可以提出问题。

(4) 自动回答问题。系统智能性的关键就在于系统能针对学生用户提出的问题自动做出回答，系统综合运用数据挖掘、中文分词、动态检索、自然语言处理、倒排索引等计算机领域的相关技术，实现自动回答。同时，智能答疑系统是否优秀的另外一个关键性指标就是自动回答的准确性、及时性、具体性、完整性，这些都是影响智能答疑系统质量的关键性要素。

(5) 教师回答问题。教师通过智能答疑系统针对学生的问题作出回答，是教师最本质的职能，且教师回答问题的答案是智能答疑系统答案库构建的数据来源，所以系统必须提供回答问题的功能。

(6) 对答案进行评价反馈。学生可以通过两种方式获得解答，首先是智能答疑系统根据学生的问题智能匹配数据库中相同或相似问题，自动给出解答，如果学生对系统自动提供的答案不满意，系统会自动将问题提交给后台，等待教师来做出解答，学生也可以对系统或教师所给出的答案做出评价。

综合以上所提出的智能答疑系统功能需求，图 3-4 为系统功能流程图。

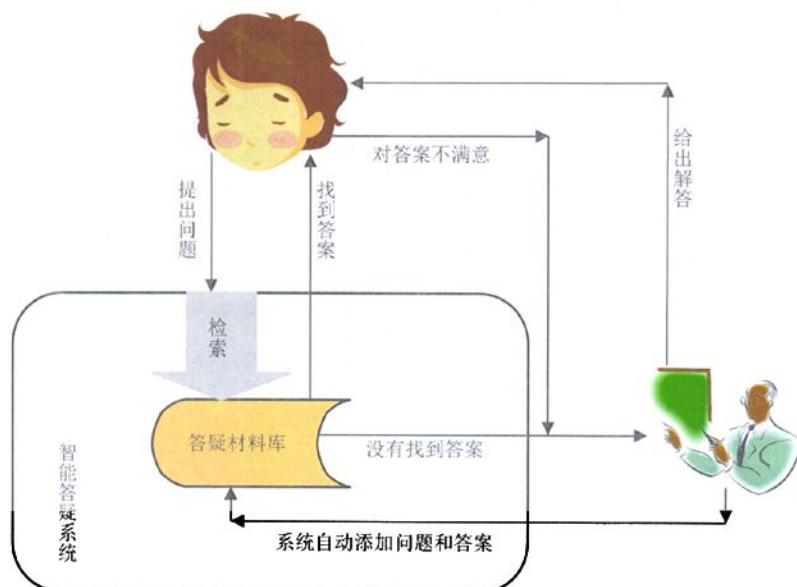


图 3-4 智能答疑系统流程图

3.4 系统非功能需求分析

远程教育模式下智能答疑系统至少需满足以下非功能性需求。

(1) 安全性。对于计算机系统来说,可靠性和安全性是极其重要的,流转于智能答疑系统中的电子数据的安全性需要得到充分的考虑。一个优秀的远程教育系统,在设计和开发时,需要懂得保护数据信息,应具有数据备份和恢复功能,所有的 Cookie 需进行多层加密、数据流转需多层校验、数据表单需多次进行传输前检查,使得数据不管处于输入、提交、传输、存储等各个阶段都应该是高度安全的。在进行数据库设计时,要让数据库结构至少满足第四范式要求,减少冗余,并且要注意防止常见的黑客攻击,比如 SQL 注入攻击等,设计合理安全的权限机制,防止非系统用户恶意发送请求、恶意刷新数据等。

(2) 易用性。不仅仅是智能答疑系统,在“服务至上”理念的高度催动下,任何一个计算机系统在研究、设计和开发过程中都要考虑到易用性要求,智能答疑系统在设计实现的过程中需考虑不同用户的不同需求,明晰框架,使得系统操作起来方便直观,且界面指示性需强,充分考虑不同用户的行为特点,不同用户都能轻易上手,确保游客用户、学生用户、教师用户都能无阻碍且得心应手地完成在线学习交流。

(3) 分布式管理。智能答疑系统需支持用户不同时间在多个不同的地点进行访问,并且快速高效,系统需建立一个核心主站以及多个枝状的地区分站,不管是游客用户、学生用户还是教师用户都可以通过主站或者分站进行注册、登录、提问、解答、浏览、搜索等操作,分布于各地区的分区定时将数据汇总至核心主站,实现实时数据交换,这样很大程度上增加了网络带宽,均衡了网络负载。

(4) 性能需求。智能答疑系统需能够经受住成两千人同时进行访问,且运转自如,并且需预留足够的存储空间来提供数据存储,系统中至少保留近十年的用户数据,超过十年的数据备份到额外硬盘,从而提高数据的读写速度,简短系统的响应时间。

3.5 系统关键技术需求分析

现如今,国内智能答疑系统发展所遇到的最大技术瓶颈就是对中文信息的处理问题,而中文信息的处理实质上就是如何将中文信息编码化、并进一步转化成系统可编译执行的语言程序的问题。中文信息处理技术是利用高速发展的计算机

技术,综合运用多领域知识、多层次算法,对中文文字信息的发音、字形、字意进行加工处理。而今中文信息处理技术日趋成熟、发展日趋稳步,对字、词、句、篇、文等内容的分析、理解、切割、生成等技术都大幅提高。现如今,随着中国进入高速发展期,中国在全世界拥有举足轻重的地位,中文信息处理技术已然发展成为自然语言信息处理技术的关键分支。

中华文字博大精深,许多汉字有多重字义、词义、句义,在不同语句中、不同环境中,相同的文字却有不同含义,正因如此,大大提高了中文信息处理技术的研究难度。虽然研究中文信息处理技术的学者数以万计、研究自然语言处理技术的学者数以百万计,但一些根深蒂固的问题始终未得到实质性解决。比如汉语语料库问题,至今为止都没有一个广受认可、放之四海而皆准的汉语语料库,更别说是带有标注的双语语料库。由于这些问题停滞不前,始终未得到解决,严重制约了答疑的精确度和准确度,致使国内智能答疑系统始终无法与一些发达国家相媲美。

所以系统需重点研究如何将中文处理技术应用到智能答疑模块中,其中应包括中文分词技术、句子相似度计算、语义相似度计算以及索引结构等。

3.6 本章小结

本章主要对智能答疑系统的用户需求、功能需求和非功能需要进行详细地分析,并阐述系统的开发目标。

第4章 系统设计

本章将对系统的设计作详细介绍,包括系统数据库设计、系统功能模块设计、系统体系结构设计三个方面。

4.1 系统体系结构

智能答疑系统注重答疑的准确性和实时性,本文设计开发的远程教育模式下智能答疑系统采用的是浏览器/服务器框架进行数据交互,基于 ASP.NET MVC 模式进行系统开发,使得多个用户可以同时通过互联网访问智能答疑系统,实时的网络在线通信大大提高了在线答疑的质量和时效性。图 4-1 为基于 ASP.NET MVC 模式设计的三层体系结构。

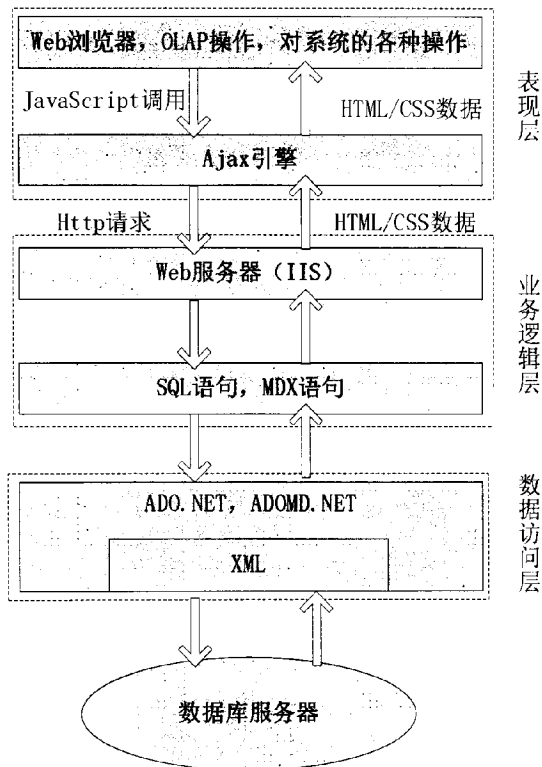


图 4-1 基于 ASP.NET MVC 模式的三层体系结构

从图 4-1 可以看出,三层体系结构分别是数据访问层、表现层、业务逻辑层。其中数据访问层是三层体系结构中的最底层,其中存储了流转于整个智能答疑系

统间的数据，这些数据包括业务数据库、用户信息、日志等，不管是文本数据、图片数据、音频数据还是视频数据等全部存储在数据访问层。表现层是与用户最直接进行交互的层级，是三层体系结构中的最上层，用户访问智能答疑系统所能看到的所有页面都属于表现层，其负责根据系统返回的数据以可视化的方式展现给用户看。业务逻辑层是三层体系结构的中间层，起桥梁作用，负责接收来自表现层的用户请求，并将其转换成数据请求后发送给数据访问层，最后并将数据访问层返回的数据再次发送给表现层，可以说所有的业务逻辑处理都在业务逻辑层。

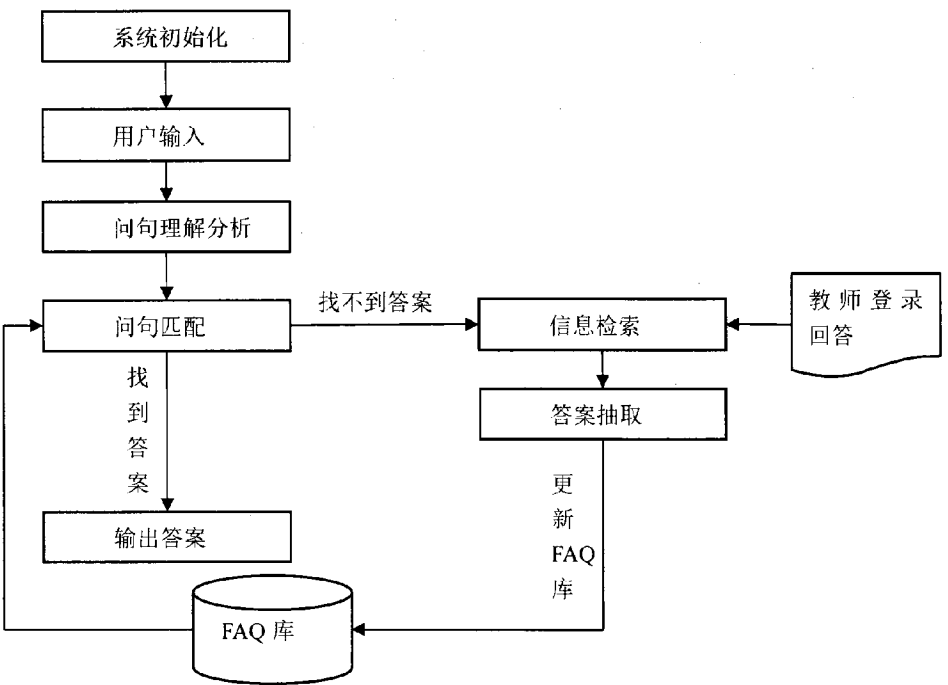


图 4-2 搜索系统结构图

通过用户需求分析可知，本系统总共有三类不同的系统用户，分别是游客用户、学生用户和教师用户。学生用户是智能答疑系统中的一类主要服务对象，是智能答疑系统最主要的用户，他们可以向系统或教师提出问题、随时查看自己所提出问题的列表、对系统的智能回答或教师的回答作出反馈评价；教师用户也是智能答疑系统中的一类主要服务对象，为学生用户解答问题，系统答案库的构建

也全依赖教师用户为学生用户源源不断地进行解答。而且,答案库的初始构建也全依赖于教师用户提供数据,以及对用户所提出问题、不满意或者低评价的回答重新解答,扩充并调整答案库资源;游客用户在访问智能答疑系统时只能简单地通过客户端浏览器浏览系统常见问题、浏览问答列表,以及搜索一些自己想要查找的信息,除此之外,游客用户不再具有其他权限。

进一步详细分析三种典型的用户类型、特点和目标,图 4-2 为智能答疑系统结构。

4.2 登录模块设计

智能答疑系统两大主要的用户群是学生用户和教师用户,两大用户具有不同的权限集。两大用户都需登录后方可访问系统后台,用户访问登录页面后,在登录页面上正确填写用户名信息、密码信息,然后提交给系统后台,系统根据数据库存储信息自动验证登录信息的正确性。如果信息错误,就返回登录页面重新填写信息;如果验证是学生用户,就跳转至学生用户主页面;如果是教师用户,就跳转至教师用户主页面。图 4-3 为系统登录模块的首页,图 4-4 为系统登录模块的功能流程图。



图 4-3 系统登录模块页面

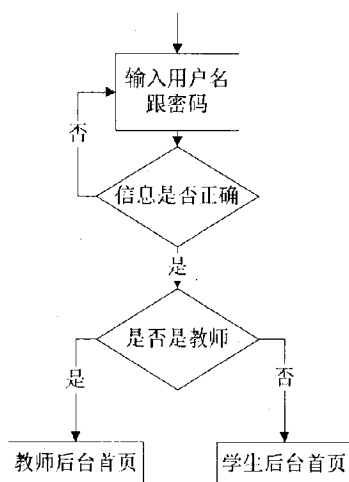


图 4-4 登录模块功能流程图

4.3 智能答疑模块设计

本文中的智能答疑系统设计了两种答疑策略，分别是自动答疑和人工答疑，以自动答疑为主、人工答疑为辅。人工答疑功能在设计和实现上相对比较简单，学生提出问题，教师根据学生提出的问题作出回答，然后系统自动将问题和答案更新至数据库。相比之下，自动答疑功能在设计和实现上相对比较复杂，在此详细坐下介绍。

4.3.1 快速中文分词机制

中文分词（Chinese Word Segmentation，简称 CWS）机制，就是指把一组特定连续的中文汉字序列按照特定的规则规范进行划分，并划分成多个单词的过程^[18-22]。当然，这些连续的中文汉字序列，可以是词语、语句、文章等等，对于语句来说，一般都有明确的分隔符进行划分，比如逗号、句号、分号、问号等等，但是词语却没有特定形式的分隔符，而不像英文语句都是以空格作为英文单词间的分隔符。所以相比于英文分词，中文分词要相对复杂的多、困难的多。

（1）汉字信息编码

每个汉字都有一个对应的汉字国标码，这是一种二进制编码，我国汉字中总共有 6763 个常用的汉字有汉字国标码，每个汉字国标码都是用一个 94*94 的方阵来表示，在方阵中，每一行叫“区”，每一列叫“位”，每个区队都从 01 到 94 进行编号，每个区块都存储一个汉字字符。

(2) 中文分词词典设计

一个优秀的中文分词机制的先决条件是要兼具快速分词、精确分词两个功能，快速分词要在保证分词效率的前提下减少多重词义语义，精确分词要解决如何对一些不常见、未录入的词汇进行识别。而两个难题需要结局，就需要设计实现一个优秀的中文分词词典，它是中文分词精确性和高效性的有力保障^[23-29]。

在此详细介绍下本智能答疑系统所设计的中文分词词典，首先是如何利用首字 Hash 法根据汉字国标码实现快速映射，如公式(4-1)。

$$offset = (C1 - 0xB0) \times 94 + (C2 - 0xA1) \quad (4-1)$$

公式(4-1)中，C1 和 C2 分别代表该汉字的汉字国标码中高低字节部分，offset 代表利用首字 Hash 法将该汉字映射到 hash 表的对应位置下标，0xB0 和 0xA1 代表常量。通过以上同时即可快速构建一张包含所有常用汉字的首字 hash 表。

完整的中文分词词典，应包括词索引表、首字 hash 表以及词典正文三层分级结构，通过三层分级结构即可快速检索中文汉字在词典中的位置，图 4-5 即为中文分词词典结构。



图 4-5 中文分词词典结构

中文分词词典设计优秀与否是影响中文分词效率的重要因素，所以本智能答疑系统采用首字 Hash 法来根据汉字国标码来加快中文分词搜索效率。

(3) 中文自动分词方法

当前，比较广为人知的中文分词算法包括以下三种，在此进行详细介绍。

① 基于字符串匹配的中文分词

基于字符串匹配的中文分词采用的是遍历查找算法，针对所要进行分词的中

文汉字序列，逐一检索中文分词词典，若在分词词典中遍历到同样的中文汉字序列，则遍历成功。

根据遍历中文汉字序列的先后顺序，可以分为正向遍历和逆向遍历两种。正向遍历是从左到右遍历中文分词词典，逆向遍历是从右到左遍历中文分词词典。

根据优先进行遍历的中文汉字序列长度是由短至长还是由长至短，可以分为最长遍历和最短遍历两种。最长遍历是先对整个中文汉字序列在中文分词词典中查找，若查找不到，就依次减短长度重新查找；最短遍历是先对中文汉字序列的单个汉字在中文分词词典中查找，若查找不到，就依次增加长度重新查找。

综合以上两种方法，我们就可知道基于字符串匹配的中文分词有以下几种常见的形式：正向最长遍历、正向最短遍历、逆向最长遍历、逆向最短遍历、双向最长遍历、双向最短遍历，且分词原理大同小异。根据不完全统计，正向最长遍历和反向最长遍历相对其他几种形式在效率上会高许多。

②基于知识理解的中文分词

所谓基于知识理解的中文分词，就是让计算机模仿人一样思考，利用到人工智能思想，通过智能理解词义、语义，在分词时消除歧义。基于知识理解的中文分词可以分为三个部分，分别是分词子系统、词义语义子系统以及总控，三者之间相互控制、相互协作。但由于汉字汉语的多样性、笼统性和复杂性，想借用人工智能思想让计算机模拟人一样思考还面临一大堆问题，想将各式各样的中文汉字让计算机理解透更是难上加难，甚至在一些发达国家，这种人工智能技术仍处于探索阶段。所以这种基于知识理解的中文分词方法尚处于试验阶段，仍不太成熟。

③基于词频度统计的中文分词

词频度也就是指一个中文词语在中文汉字序列中出现的次数，而中文词语又是由中文汉字组成的，所以相邻的中文汉字在中文汉字序列中出现的越多，这些相邻的中文汉字组成中文词语的概率就愈大，这就称为中文汉字的紧密程度。当这种紧密度值超过某一个提前设定的阈值时，这些相邻的中文汉字就能组成一个中文词语。基于词频度统计的分词方法最大的优点就是无需提前构建中文分词词典，也无需担心中文词语的词义语义问题，只需计算词频度即可。但是该方法最大的缺点就是分词准确性问题，有时候词频度高的不一定能组成中文词语，所以

对语料库的规模要求很高，计算难度较大。

综上所述，本文中的智能答疑系统综合考虑计算复杂度、分词精确度等问题后，决定采用基于字符串匹配的中文分词方法。

(4) 停止词处理

所谓停止词，是指在中文汉字序列中无过多词义的词语，这些词语是无需进行索引的，比如“是”、“什么”、“的”等，这些词语不在少数，可能会浪费许多索引空间，降低索引效率，所以特意设计一张停止词表进行存储。

虽然说停止词绝大多数情况下是无意义的，但是对于停止词的处理，又不能直接提前将中文汉字序列中所有的停止词去除，因为不是所有的停止词都是没有意义的，比如“的士”中的“的”在词中是有意义的，如果提前去除是错误的。本智能答疑系统在对中文汉字序列分词后，再对停止词进行判断，增加了分词的准确度。

4.3.2 基于向量空间模型的句子相似度计算

(1) 向量空间模型

上个世纪中期，Salton 等人提出向量空间模型，并将其用于 SMART 系统，只是 SMART 系统称为当时首个使用向量空间模型的文本检索系统。向量空间模型通过向量运算来计算中文语句间的语义相似度，简化文本处理过程，从而用于实现索引、相关性评估、信息提取、信息过滤等操作。

对于文件来说，它由一系列中文词语组成，如果将在文件中出现过一次的中文词语看成一元向量，词频度看成向量长度，那么单个文件可以看作由多个索引词组成的多元向量空间，而两个文件的相似度就是两个多元向量空间的距离问题。而计算文件相似度十分复杂，难度也大，所以一般都将多元向量空间进行降维计算，常用的降维方法有标点符号、停止词等等，将其简化为句子相似度计算，甚至是词语相似度计算。

(2) TF-IDF 方法

Term Frequency - Inverse Document Frequency，简称 TF-IDF，是一种常见的用于统计某个词语对某个文档或者某个中文汉字序列的重要程度的加权统计方法，常用于搜索引擎中进行资讯勘探、资讯检索。它利用中文汉字或中文词语在文件中的出现频率来评估其重要程度，出现频率越高，表示该中文汉字或中文词

语越重要。此外, TF-IDF 还综合多个文件来评估该中文汉字或中文词语的重要程度, 如果该中文汉字或中文词语在该文件中出现频繁, 且少见与其他文件, 那么该中文汉字或中文词语即可作为该文件的特征词或关键词。

在向量空间模型中, 假设权重 $W_{i,j}$ 代表问句 j 中第 i 个词的权重。 $W_{i,q}$ 代表查询 q 中第 i 个词的权重。问句 j 和查询 q 可以分别表示为。

$$\vec{d_j} = (w_{1,j}, w_{2,j}, \dots, w_{i,j}) \quad (4-2)$$

$$\vec{q} = (w_{1,q}, w_{2,q}, \dots, w_{i,q}) \quad (4-3)$$

如果有以上两组向量, 可以采用两个向量夹角的余弦值作为两个向量的相似度。

$$Sim(\vec{d_j}, \vec{q}) = \frac{\vec{d_j} \cdot \vec{q}}{|\vec{d_j}| \times |\vec{q}|} = \frac{\sum_{i=1}^n (w_{i,j} \times w_{i,q})}{\sqrt{\sum_{i=1}^n w_{i,j}^2} \times \sqrt{\sum_{i=1}^n w_{i,q}^2}} \quad (4-4)$$

其中的 $w_{i,j}$ 可以表示如下。

$$w_{i,j} = f_{i,j} \times \log \frac{N}{n_i} \quad (4-5)$$

等式中的 $f_{i,j}$ 表示如下。

$$f_{i,j} = \frac{freq_{i,j}}{\max_i freq_{i,j}} \quad (4-6)$$

其中 $freq_{i,j}$ 表示第 i 个词在第 j 文本中的频率。

$w_{i,j}$ 中的 $\log(N/n_i)$ 表示为。

$$idf_i = \log \frac{N}{n_i} \quad (4-7)$$

其中 $w_{i,j}$ 便是 TF-IDF 公式, 这里 N 是文本的总数量, n_i 是包含索引词 k_i 的文本数量。

综合以上公式, 可以得到 TF-IDF 法的公式为。

$$W = TF \times IDF \quad (4-8)$$

例如, 假设本智能答疑库中某个句子所包含的中文词语可以表示为 (w_1, w_2, \dots, w_n) , 那么每一个中文语句可以进一步表示为一个维度是 n 维的空间向量, 格式为 $T = \langle T_1, T_2, \dots, T_n \rangle$, 而其中每一个 $T_i (1 \leq i \leq n)$ 的算法公式如下。

$$T_i = TF \times IDF = n \times \log \frac{M}{m} \quad (4-9)$$

其中 n 表示词 w_i 在句子中出现的次数, m 表示包含词 w_i 的句子个数, M 为句子的总数。

可以用与上面同样方法计算问句的 n 维向量 $T' = \langle T'_1, T'_2, \dots, T'_n \rangle$ 。得到目标问句的 T 和 T' 后, 两个语句的相似度即可通过向量余弦值进行计算。

$$Sim(T, T') = \frac{\sum_{i=1}^n (T_i \times T'_i)}{\sqrt{\sum_{i=1}^n T_i^2} \times \sqrt{\sum_{i=1}^n T'^2_i}} \quad (4-10)$$

4.3.3 基于《同义词词林》的语义相似度计算

语义相似度计算是中文汉字处理中难度较大的一项内容, 其包括相似词语定义、相似词语关联等内容, 在中文自动翻译以及搜索引擎检索领域都有极为广泛的应用。语义相似度计算应用较为普遍广泛的方式是基于语义词典的语义相似度计算, 比如国内的《同义词词林》、《中文概念词典》、《知网》等, 国外的 FrameNet、MindNet、WordNet 等。本智能答疑系统就是采用《同义词词林》来计算词语间的语义相似度。

(1) 《同义词词林》简介

《同义词词林》是由梅家驹等人于 1983 年撰写的, 包含了许多中文词语的同义词, 后由不同的科研机构、不同的学者对《同义词词林》的同义词数量进行更新, 更有科研机构推出扩展版, 收录了几万条的中文词语及其同义词, 并根据词语同义性进行排序。同义词概念在数据挖掘、数据提取、数据检索等领域得到广泛应用, 尤其近几年大数据概念盛行, 使得中文词语同义词越来越受到关注。

(2) 词与词之间语义相似度计算

《同义词词林》是基于汉字的使用原则和使用特点来进行词语的语义分类的, 一般情况下可以划分为 1428 个小类、94 个中类以及 12 个大类, 图 4-6 为语义结构图。

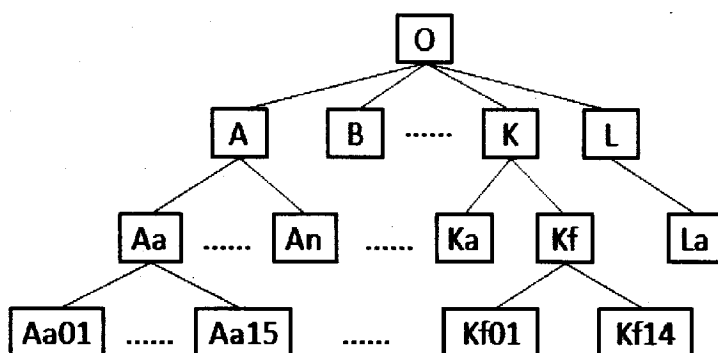


图 4-6 语义结构

正常情况下，常使用两个词语在语义结构上的“距离”来计算词语间的语义相似度，基于此方法可以将《同义词词林》进行分类。

假设 A、B 两个词语的语义编码分别用 a、b 表示，那么 A、B 两个词语间的语义距离即可通过以下公式进行计算。

$$Dist(A, B) = \min_{a \in P, b \in Q} dist(a, b) \quad (4-11)$$

其中词语 A 的语义集合用字母 P 表示，词语 B 的语义集合用字母 Q 表示，词语 A 和 B 之间的语义距离即可表示为。

$$dist(a, b) = 2 \times (4 - n) \quad (4-12)$$

其中 n 表示从第 n 类开始，词语 B 与词语 A 之间的直接语义编码不一样，这样假如 n 的值为 0，就代表词语 A 和 B 的语义完全相同。可用以下公式进行计算。

$$Similar(A, B) = \frac{1}{Dist(A, B)} \quad (4-13)$$

从公式中可以看出，词语与词语之间的语义相似度跟语义距离是成负相关的。

(3) 句子与句子间的语义相似度计算

语义相似度是基于词语间的语义相似度来计算的，假设 A、B 两个句子分别包含词组 A_1 、 A_2 、 \dots 、 A_m 和词组 B_1 、 B_2 、 \dots 、 B_n ， $s(A_i, B_j)$ 表示 $B_j (1 \leq j \leq n)$ 和 $A_i (1 \leq i \leq m)$ 的词语语义相似度，即可得到一个大小为 $m \times n$ 的矩阵，格式为。

$$M(A, B) = \begin{bmatrix} s(A_1, B_1), s(A_1, B_2), \dots, s(A_1, B_n) \\ \dots\dots\dots \\ s(A_m, B_1), s(A_m, B_2), \dots, s(A_m, B_n) \end{bmatrix} \quad (4-14)$$

而 $s(A, B)$ 即可表示为。

$$s(A, B) = \sum_{i=1}^m \frac{\max(s(A_i, B_1), s(A_i, B_2), \dots, s(A_i, B_n))}{m} \quad (4-15)$$

接着对通过公式 (4-10) 和 (4-15) 计算的相似度值进行加权平均, 得到两句子最终的相似度值为。

$$m = \alpha T + \beta S \quad (4-16)$$

其中 T 和 S 分别表示 TF-IDF 算出的相似度和用语义计算得出的相似度, α , β 分别表示相应的权重因子。

4.3.4 索引结构

索引结构有正排索引结构和倒排索引结构两种, 正排索引结构是将中文语句划分成多个关键词后, 将关键词根据该中文语句进行索引, 格式如下。

Question1 \rightarrow KeyWord1, KeyWord2, KeyWord3, KeyWord4 $\dots\dots$

也就是说, 每个中文语句都可以表示成一连串关键词的集合, 而且不管是语句还是关键词都有指定的 ID 号与其对应, 这样就可以简化字符串的存储, 只需对 ID 号进行存储, 图 4-7 为正排索引结构。

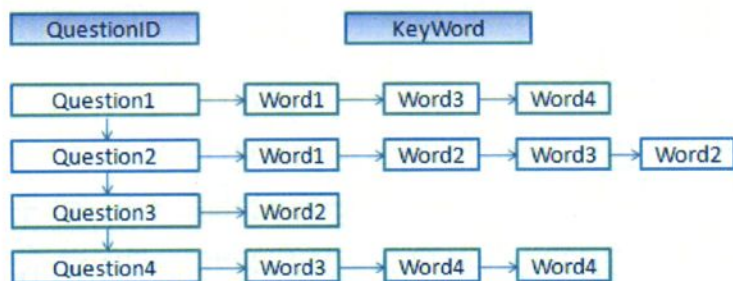


图 4-7 正排索引结构

而倒排索引结构与正排索引结构相反, 是将中文语句划分成多个关键词后, 将中文语句根据关键词进行索引, 本智能答疑系统就是采用这种结构进行索引,

格式如下。

Key Word1→Question1, Question2, Question3, Question4

图 4-8 为倒排索引结构。

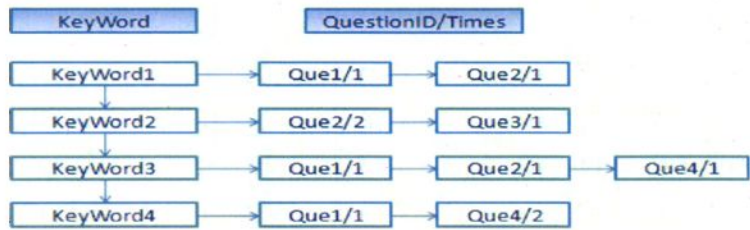


图 4-8 倒排索引结构

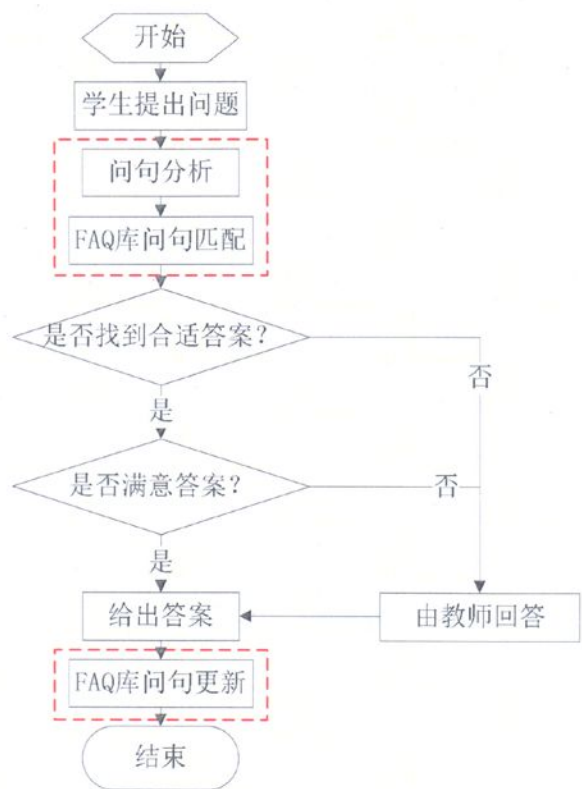


图 4-9 自动答疑模块工作流程图

综上所述即为自动答疑模块中所用到的关键技术，图 4-9 为自动答疑模块的工作

流程图，从图中可以看出，自动答疑模块至少包含了 FAQ 库问句更新子模块、FAQ 库问句匹配子模块以及问句分析子模块，也就是图 4-9 中红色虚线框框住的部分。

(1) 问句分析子模块。问句分析子模块主要负责在学生用户提交问题后，利用中文分词、词性标注等技术，辨别问题所属类型、提取问题关键词、利用词义相似度扩展同义词并确定问题所属的知识单元，从而完成对问题语义的理解，并以中间语言的形式来表示问句分析的结果。

(2) FAQ 库问句匹配子模块。学生用户在提交问题后，智能答疑系统会首先将其和 FAQ 库中的问题进行对比分析，查找和 FAQ 库中相似度最高的问题，当相似度大于某一个已设定的阈值时，就为学生用户返回所需答案，否则就将问题提交给后台等待教师解答。FAQ 库问句匹配子模块能够大大提高答疑的效率，无需经过复杂的流程，无需经过复杂的计算，有效地提高了智能自动答疑的效率和准确率。

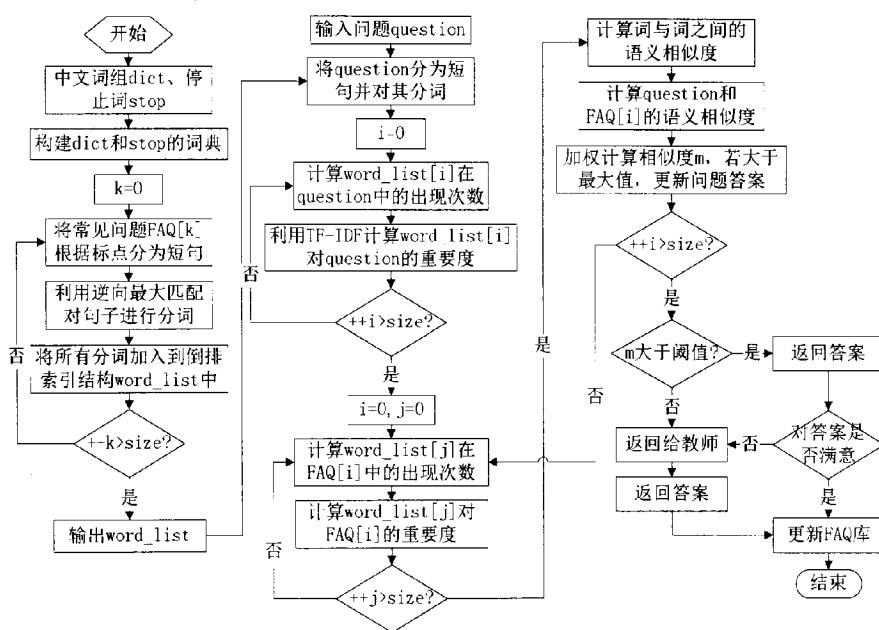


图 4-10 智能答疑模块算法流程图

(3) FAQ 问句更新子模块。用户在使用智能答疑系统时，有两种情况会触发 FAQ 问句更新子模块，一种是在智能答疑系统自动解答学生用户提出的问题

时,一种是当学生用户对系统自动解答的答案不满意,由教师进行补充回答时,这两种情况下都会触发FAQ库的自动更新。

图4-10为该智能答疑模块的算法流程图。

4.4 数据库设计

系统数据库的设计是整个系统设计开发过程的关键性部分,其设计结构的优劣性直接影响了数据是否安全、性能是否高效,所以在进行数据库结构设计时,需数据索引结构、数据检索效率、数据规范化等问题。在设计智能答疑系统数据库时,充分考虑以上问题,让数据库设计满足第四范式要求,避免出现过多的数据冗余,大大提高了数据交互的效率。

4.4.1 逻辑关系图(E-R图)

详细分析系统功能需求后,设计了本智能答疑系统的数据库,图4-11为系统数据库的逻辑结构图。

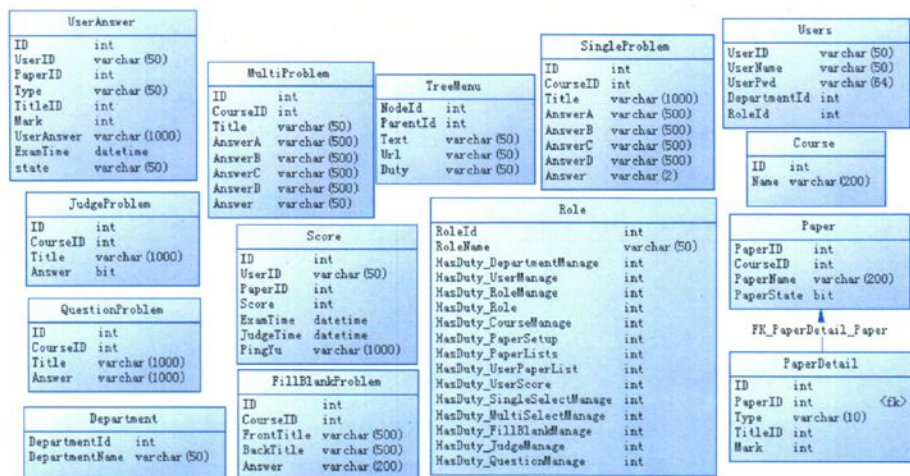


图4-11 数据库E-R图

4.4.2 数据库表结构设计

以下介绍几张比较关键性的数据表。

1、Dic 表

表4-1为Dic表,主要用来保存中文词组。

表 4-1 Dic 表

字段	说明	Null	类型
word	中文词组	否	nvarchar(50)

2、Stop 表

表 4-2 为 Stop 表，主要用来保存中文分词停止词。

表 4-2 Stop 表

字段	说明	Null	类型
word	中文分词标志词	否	nvarchar(50)

3、user 表

表 4-3 为 user 表，主要用来保存用户的基本信息。

表 4-3 user 表保存用户的基本信息

字段	说明	Null	类型
username	用户登录名	否	nvarchar(50)
usertype	用户类型	否	int
password	用户登录密码	否	nvarchar(50)

4、Similar 表

表 4-4 为 Similar 表，主要用来保存同义词编码信息。

表 4-4 Similar 表保存同义词编码

字段	说明	Null	类型
id	主键 id	否	int
WORD	中文单词	否	nvarchar(50)
ATTR	同义词编码	否	nvarchar(50)

5、FAQ 表

表 4-5 为 FAQ 表，主要用来保存常用问答库信息。

表 4-5 FAQ 表

字段	说明	Null	类型
fid	主键 id	否	int
ANS	答案	是	nvarchar(4000)
QUE	问题	是	nvarchar(4000)

6、Question 表

表 4-6 为 Question 表，主要用来保存问题库信息。

表 4-6 Question 表

字段	说明	Null	类型
Q_ID	主键 id	否	int
A_Time	回答时间	是	datetime
A_User	回答者	是	nvarchar(50)
A_Content	教师回答内容	是	nvarchar(4000)
Q_OK	是否满意	否	tinyint
Q_Reply	自动回答的答案	否	nvarchar(4000)
Q_Time	提问时间	否	datetime
Q_User	提问者	否	nvarchar(50)
Q_Content	问题内容	否	nvarchar(4000)
Q_Title	问题标题	否	nvarchar(100)

4.5 本章小结

本章对系统的设计作详细介绍，包括系统数据库设计、系统功能模块设计、系统总体结构设计三个方面。

第5章 系统实现与测试

本章是系统实现与测试部分，主要对系统的实现过程进行详细介绍，并对系统进行全面地功能测试、性能测试等。

5.1 系统三层体系结构的实现

5.1.1 Model 部分的实现

本智能答疑系统是基于 ASP.NET MVC 模式设计开发的，Model 部分负责存储本系统的所有应用数据，可以说，Model 部分相当于数据存储、处理的仓库。使用 ASP.NET MVC 框架使得可以使用任何一种数据库进行开发，且无需顾虑后期数据库的更换问题，本智能答疑系统选用微软公司开发的 SQL SERVER 2008 作为系统数据库，并根据之前介绍的数据库设计结构来设计数据库表。

可以用来构建对象关系映射的框架有很多，本智能答疑系统选用 LINQ to SQL 框架来处理 .NET 类对象和系统数据库之间的一一映射关系，让用户忽略底层数据库详细的持久化操作过程，转化成简便的对象类操作。

首先在 Visual Studio 2010 开发环境上新建名为“DSReply”的 Web 项目，然后在 DSReply 项目中新建名为“DSReply.dbml”的 Linq To Sql 类，根据数据库表结构编写类对象，或者简便地将数据库表拖拽进 VS2010，即可自动新建与数据库表相对应的类对象。

本智能答疑系统主要涉及六张数据库表，利用 Linq To Sql 技术实现类对象和数据库表之间的一一映射关系很简单，只需添加以下两行代码即可。

```
[global::System.Data.Linq.Mapping.DatabaseAttribute(Name="DSReply")]  
public partial class DSReplyDataContext : System.Data.Linq.DataContext
```

而数据库中所有的数据表都对应生成了与其对应的类对象，而每个类对象中的成员属性即与数据表中列形成一一对应。

这样就完成了数据表到 .NET 类对象的意义映射，可以直接编写 LINQ 语句来直接操作数据库了。例如利用以下 LINQ 语句即可返回 user 数据表中的全部数据。

```
user_list = db.users.ToList();
```

又例如利用以下 LINQ 语句即可返回 question 数据表中 ID 属性值为 id 的问题。

```
var q = from s in db.Questions
        where s.Q_ID == id
        select s;
ques.question = q.ToList();
```

5.1.2 View 部分的实现

在 VS2010 中, View 部分页面类型可以是 master、aspx 或 ascx 等, 本智能答疑系统主要采用 aspx 和 master 文件, 其中 master 文件代码如下所示。

```
<html xmlns="http://www.w3.org/1999/xhtml">
<head runat="server">
    <title>智能答疑系统</title>
    <asp:ContentPlaceholder ID="TitleContent" runat="server" />
    <link href="../../../Content/main.css" rel="stylesheet" type="text/css" />
    <link href="../../../Content/common.css" rel="stylesheet" type="text/css" />
</head>
<body id="ct100_body">
    <div id="container">
        <!-- Banner Begin-->
        .....
        <!-- Banner End-->
        <div id="menu">
        </div>
        <div id="main">
            <asp:ContentPlaceholder ID="MainContent" runat="server" />
        </div>
        <div class="clear">
        </div>
    </div>
    <div id="footer">
    </div>
</body>
</html>
```

而其他页面只需在母版页的基础上,修改 ContentPlaceHolder 控件,再根据各自页面的详细内容进行编写即可。例如 FAQ.aspx 的页面代码如下所示。

```
<%@ Page Title="" Language="C#" MasterPageFile="~/Views/Shared/Site.Master"
Inherits="System.Web.Mvc.ViewPage<DSReply.Controllers.HomeController+Faq>" %>
<asp:Content ID="Content1" ContentPlaceHolderID="TitleContent" runat="server">
    常见问题列表
</asp:Content>
<asp:Content ID="Content2" ContentPlaceHolderID="MainContent" runat="server">
    <div id="mainLeft">此处省略代码</div>
    <div id="mainRight">此处省略代码</div>
</asp:Content>
```

5.1.3 Controller 部分的实现

(1) 路由映射机制

Controller 部分是介于 View 部分和 Model 部分之间,主要是用于接受并处理用户请求,而 ASP.NET 独有的路由映射机制就是负责在 Controller 部分拦截并解析用户请求地址,并根据地址调用相应的 Controller 访问相应的 Model 来响应用户请求,并将响应结果返回给 View。所有路由映射机制是 ASP.NET 中一项非常关键性的技术,而编程人员只需根据路由映射规则编写相应的 URLs 格式即可。

为了加载 Controller 部分的路由映射机制,需首先在项目“DSReply”中添加以下一小段代码。

```
<add assembly="System.Web.Routing, Version=4.0.0.0, Culture=neutral,
PublicKeyToken=31BF3856AD364E35"/>
```

然后新建类对象,并在里面设定路由映射规则,该路由映射规则可以采用开发平台自带的默认规则,也可以根据自己需要设定指定规则。本智能答疑系统采用的是 ASP.NET 平台自带的路由映射规则,代码如下所示。

```
public static void RegisterRoutes(RouteCollection routes)
{
    routes.IgnoreRoute("{resource}.axd/{*pathInfo}");
    routes.MapRoute(
```

```
"Default", // Route name
"{controller}/{action}/{id}", // URL with parameters
new { controller = "Home", action = "Index", id = "" } // Parameter defaults
);
}
```

至此，就成功地在智能答疑系统中添加了 ASP.NET 平台的路由映射机制，所采用的是默认的映射规则。比如当通过客户端浏览器访问 <http://localhost/Home/Index>，路由映射机制就会自动根据地址解析出 action 名是 Index、控制器名为 Home，然后根据解析出的控制器和 action，调用名为 HomeController 类中的 Index 方法，最后并将运行的结果和参数添加到名为 Home/Index.aspx 的页面然后返回给客户端。

2、制定 Controller

在引入路由映射机制并制定路由映射规则后，就可以构建 Controller 了，本智能答疑系统总共构建了 TeacherController、StudentController 和 HomeController 三个控制器类。其中 TeacherController 负责教师用户的业务处理，StudentController 负责学生用户的业务处理，HomeController 负责智能答疑系统的登录控制、首页展示等业务处理。

Controller 类中所有公用方法，也就是 public 方法都被当做是 Action 方法，而且方法名就是 Action 名，如前面举的例子，当访问 Home/Index 时，其中 Index 既是 Action 名也是 Action 方法的方法名，而且每个 Action 方法通常都会返回 ActionResult 对象。

至此，本系统的 ASP.NET MVC 结构就已经基本构建好了，图 5-1 为系统目录结构图。

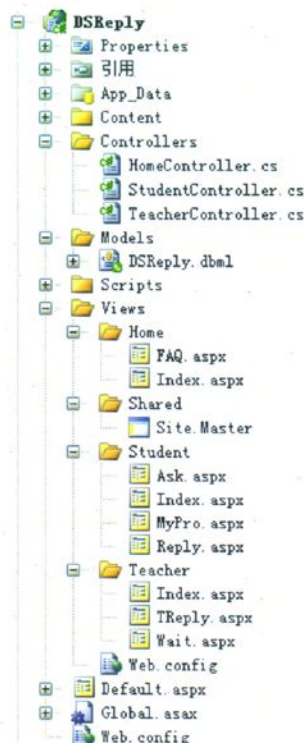


图 5-1 系统目录结构

5.2 登录模块实现

智能答疑系统其中一块非常重要的部分就是安全性，本智能答疑系统采用的是自定义的表单验证方式来验证用户的权限和信息，其中整个登录功能分为表单的提交和验证两部分，其中表单设计的代码如下。

```
<%if (Session["username"] == null)
{
    %>
    <% using (Html.BeginForm("Login", "Home"))
    {
        %>
        用户名
        <%= Html.TextBox("username", "", new { @size = "12" }) %>
        密码
        <%= Html.Password("password", "", new { @size = "12" }) %>
        <input id="button" class="an" type="submit" value="登录" name="button" />
    }
    %>
```

编写并运行以上代码后也可见表单页面，系统用户在访问表单中填写用户名、密码等信息，填写结束后提交表单至名为 HomeController 的类中的 Login 方

法，并在 Login 方法中编写以下代码进行登录验证。

```
for (int i = 0; i < user_list.Count; i++) {
    if (form["username"] == user_list[i].username && form["password"] ==
user_list[i].password) {
        Session["username"] = form["username"];
        Session["usertype"] = user_list[i].usertype.ToString();
        if (user_list[i].usertype == 0) return RedirectToAction("../Teacher/Index");
        else return RedirectToAction("../Student/Index");
    }
}
```

根据以上代码可以看出，智能答疑系统首先验证用户输入的信息是否正确，如果信息不正确，系统就自动跳转回登录页面让用户重新输入。如果数据的用户名和密码信息全部正确，系统再进一步判断登录系统的用户是教师用户还是学生用户，若是教师用户，就自动跳转至教师后台页面，若是学生用户，就自动跳转至学生后台页面。其中图 5-2 和图 5-3 分别为学生用户和教师用户的系统后台首页图。

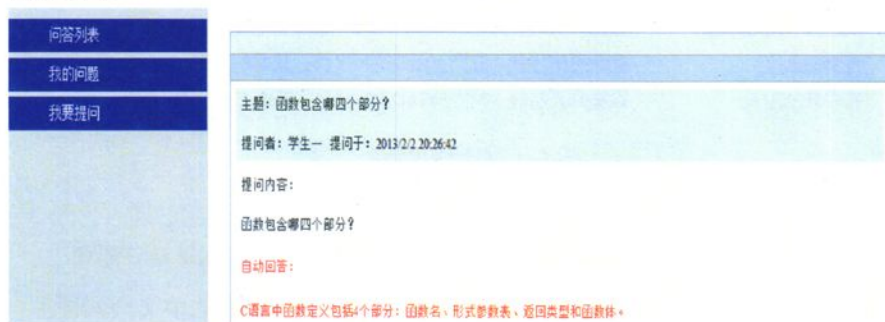


图 5-2 学生后台首页

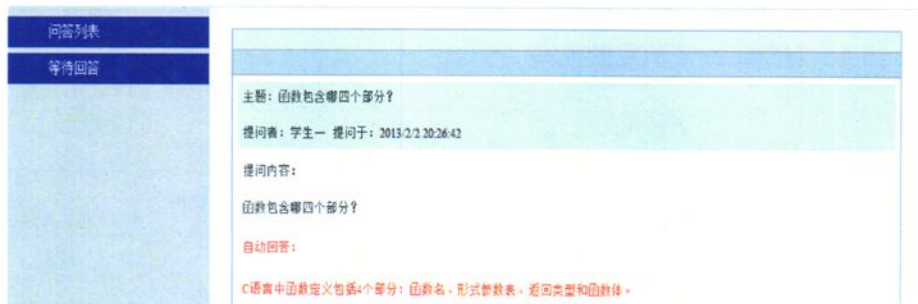


图 5-3 教师后台首页

5.3 智能答疑模块实现

整个智能答疑系统能够畅通无阻运行的重要数据基础是构建强大的答疑库，而且答疑库在构建的过程中必须保证其数据规模及完备性，其是智能答疑准确性和智能性的先决条件。由于在智能答疑系统构建初期，缺乏师生交互数据，所以答疑库的初始化工作就由来完成，图 5-4 即为构建的初始答疑库数据。

QUE	ANS	fid
合并排序的基本思想	合并排序算法是用分治策略实现对n个元素进行排序...	25
线性时间排序算法有哪些	计数排序和桶排序。	26
计数排序的基本思想	计数排序算法的基本思想是对每一个输入元素x，确...	27
桶排序的基本思想	桶排序算法的基本思想是：设置若干个桶，将键值...	28
树的定义	树是由一个集合以及在该集合上定义的一种层次关...	29
度的定义	一个结点的儿子结点数称为该结点的度。一棵树...	30
叶结点与分支结点	树中度为零的结点称为叶结点或终端结点。树中度...	31
结点的高度和树的高度	树中一个结点的高度是指从该结点到各叶结点的最...	32
树的遍历	树的遍历是树的一种重要的运算。所谓遍历是指对...	33
表示树的几种基本方法	父节点数组表示法、儿子链表表示法、左儿子右兄...	34
实现二叉树的方法	二叉树的顺序存储结构、二叉树的结点数表示法、...	35
线索二叉树的定义	用指针实现二叉树时，每个结点只有指向其左、右...	36
集合的定义	集合是由元素（成员）组成的一个类。集合的成员...	37
符号表的定义	以集合为基础，并支持SetMember、SetInsert和S...	38

图 5-4 初始答疑库

中文信息处理是智能答疑模块的重点，其所需用到的技术和算法，都在上一章中进行了介绍，主要包括中文分词词典设计、基于字符串匹配的中文分词技术、倒排索引结构、基于向量空间模型 TF-IDF 方法来计算句子相似度、采用语义距离计算语义相似度、Similar 数据表的设计等。

图 5-5 为 Similar 表的数据格式。

id	ATTR	WORD
4	Al04	阿斗
5	Br14	阿片
6	Ab02	阿妮
7	Ah04	阿母
8	Ah02	阿爷
9	Ah04	阿妈
10	Ah05	阿伯
11	Ah10	阿妹
12	Ah09	阿弟
13	Ah04	阿爸
14	Ah05	阿叔
15	Ah07	阿姑
16	Ah13	阿姑
17	Ah10	阿姐
18	Ah10	阿妹
19	Ae13	阿姨
20	AED2	阿姨

图 5-5 Similar 表数据

这样就可利用 TF-IDF 计算出句子相似度、利用语义距离计算语义相似度，并利用公式（4-16）对两个相似度进行加权求和，并计算 FAQ 库中各个问句的综合相似度值 m 以及最大的 m 值 $\max m$ 。如果 $\max m$ 大于或等于我们预设的阈值 M ，就可以认为学生用户提问的问题和 FAQ 库中的该问句是相似度最高的问题；如果 $\max m$ 小于我们预设的阈值 M ，就可以认为学生用户提问的问题和 FAQ 库中的所有问句都不匹配，那么系统就会自动将该问题提交给相应的教师用户等待解答。在本智能答疑系统中将 α 、 β 以及 M 分别取值 0.6、0.4、0.6。

由于 FAQ 库数据量较为庞大，且许多情况下学生用户提问的问题都不能在 FAQ 库中找到相似性问句，为了提高效率，通常会先缩小检索范围，从 FAQ 库中筛选出部分数据。本智能答疑系统中，首先会进行中文分词并提取关键词，然后筛选出 FAQ 库中至少与其有一个相同关键词的问句作为候选集合，然后再在这个候选集合中进行以上计算。

本智能答疑系统在进行关键词索引时采用的是倒排索引结构，在对用户所提交问句的关键词进行索引时，首先采用二分查找法查找关键词在倒排索引表中的位置，然后从头遍历这个关键词所指向的 FAQ 库中的问句，就可以轻易地知道在 FAQ 库中哪些问句包含这个关键词。通过上述的方式，遍历用户提交问句中的所有关键词，即可进一步寻找出相似度最高的问句。

其中涉及到主要代码如下。

1、分词词典设计

```
int tot = word.Count;
for (int i = 0; i < tot; i++)
{
    char[] c = new char[1]; c[0] = word[i][0];
    byte[] b = Encoding.Default.GetBytes(c); int id = (b[0] - 176) * 94 + (b[1] - 161);
    if (Index[id].cnt == 0) {
        Index[id].ch = c[0]; Index[id].cnt++; Index[id].first = i;
    }
    else {
        Index[id].cnt++;
    }
}
}
```

2、逆向最大匹配分词

```
//逆向最大匹配分词
private ArrayList post_max(string text, Dictionary dict)
{
    ArrayList word = new ArrayList();
    int len = text.Length, endPos = len - 1, cnt, step; string unknown = ""; char[] c = new char[1];
    while (endPos >= 0) {
        if (Utility.isDouble(text[endPos]) == false) {
            if (unknown.Length != 0) { char[] a = unknown.ToCharArray(); Array.Reverse(a);
            word.Add(new string(a)); unknown = ""; }
            cnt = 0;
            while (endPos >= 0 && Utility.isDouble(text[endPos]) == false) { cnt++; endPos--; }
            endPos++; string cur = text.Substring(endPos, cnt); cur = cur.ToLower();
            word.Add(cur); endPos--; if (endPos < 0) break;
        }
        for (step = Utility.WORD_MAX; step >= 1; step--) {
            if (endPos - step + 1 < 0) continue; string cur = text.Substring(endPos - step + 1, step);
            if (dict.find(cur) == true) {
                if (unknown.Length != 0) { char[] a = unknown.ToCharArray(); Array.Reverse(a); word.Add(new string(a)); unknown = ""; }
                word.Add(cur); endPos = step; break;
            }
        }
        if (step < 1) { unknown += text.Substring(endPos, 1); endPos--; if (endPos < 0) break; }
    }
}
```

```

    if (unknown.Length != 0) { char[] a = unknown.ToCharArray(); Array.Reverse(a);
    word.Add(new string(a)); }
    return word;
}

```

3、倒排索引结构

```

public class WORD_NODE
{
    public int index;
    public string word; //词
    public int cnt;      //词在问题库中出现的次数
    public WORD_NODE() { }
    public WORD_NODE(string str, int index) { this.cnt = 1; this.word = str; this.index =
index; }
}

```

4、TF-IDF 方法

```

//计算TF-IDF  $n \cdot \log(M/m)$ , num_in_que表示词语在用户问题中出现的次数, tot_que表示FAQ库中的
问题数, m_s表示包含该词语的FAQ问题数
public static double getT(int num_in_que, int tot_que, int que_hav_num)
{
    if (que_hav_num == 0) return 0.0;
    return (double)num_in_que * Math.Log((double)tot_que / que_hav_num, 2);
}

```

5、句子语义相似度计算

```

// 计算两个句子语义相似度
public static double GetSimilar(ArrayList a, ArrayList b)
{
    DSReplyDataContext data = new DSReplyDataContext(); ArrayList a_attr = new ArrayList();
    ArrayList b_attr = new ArrayList();
    for (int i = 0; i < a.Count; i++) { ArrayList t_attr = new ArrayList();
        var q = from s in data.Similar where s.WORD == a[i].ToString() select s.ATTR;
        IList<string> word = q.ToList(); if (word.Count > 0) t_attr.Add(word[0]);
        a_attr.Add(t_attr);
    }
}

```

```

for (int i = 0; i < b.Count; i++){ ArrayList t_attr = new ArrayList();
    var q = from s in data.Similars where s.WORD == b[i].ToString() select s.ATTR;
    IList<string> word = q.ToList(); if (word.Count>0) t_attr.Add(word[0]);
b_attr.Add(t_attr);
}
int[] flag = new int[a.Count]; Array.Clear(flag, 0, a.Count);
for (int i = 0; i < a.Count; i++){ int k; for (k = 0; k < b.Count; k++) if
(a[i].Equals(b[k])) {break;}
    if (k < b.Count) flag[i] = 1;
}
double cnt = 0.0;
for (int i = 0; i < a_attr.Count; i++){ if (flag[i] == 1){cnt += 1.0; continue;}
    if (((ArrayList)a_attr[i]).Count == 0) continue;int dis = -1;
    for (int j = 0; j < ((ArrayList)a_attr[i]).Count; j++){
        for (int ii = 0; ii < b_attr.Count; ii++){ if (((ArrayList)b_attr[ii]).Count
== 0) continue;
            for (int jj = 0; jj < ((ArrayList)b_attr[ii]).Count; jj++){
                int d = dist(((ArrayList)a_attr[i])[j].ToString(),
((ArrayList)b_attr[ii])[jj].ToString());
                if (d > dis) dis = d;
            }
        }
    }
    if (dis == -1) cnt += 0.0; else if (dis == 4) cnt += 1.0; else cnt += 1.0 / (double) (2.0
* (4 - dis));
}
return cnt / (double)a.Count;
}

```

6、加权计算句子相似度并与阈值比较

```

for (int i = 0; i < faq_list.Count; i++){vector value = new vector();
    for (int j = 0; j < word_list.Count; j++){
        int n = getANSN(i, (ArrayList)index list[(((WORD_NODE)word_list[j]).index]);
        int m = ((WORD_NODE)word_list[j]).cnt; value.v.Add(Utility.getT(n, faq_list.Count,
m));
    }
    cross_value = 0.6 * value_s.cross(value) + 0.4 * Utility.GetSimilar(res_s,
(ArrayList)faq_split[i]);
    if (cross_value > ans_value){ans_value = cross_value; ans_id = i;}//计算maxm
}
string reply = "无法回答，等待教师进行答复！";
if (ans_value < Utility.value) { Session["Answer"] = reply; return View(); }

```

```
reply = faq_list[ans_id].ANS; Session["Answer"] = reply;
```

7、FAQ 库更新

```
if (action == "满意")
{
    ques.Q_OK = 1;
    FAQ faq = new FAQ();
    faq.fid = db.FAQs.ToList().Count + 1;
    faq.QUE = ques.Q_Content;
    faq.ANS = ques.Q_Reply;
    db.FAQs.InsertOnSubmit(faq);
    db.SubmitChanges();
}
```

此外，智能答疑模块主要的系统界面如下所示，图 5-6 为学生用户进行提问的页面；图 5-7 为智能答疑系统系统对学生用户所提问题进行自动回答的页面；如果学生用户对自动回答的结果满意，智能答疑系统会将该问题加入 FAQ 库，如图 5-8 所示；否则，智能答疑系统会将学生用户所提问题提交到教师用户后台等待其解答，如图 5-9 所示；教师用户回答问题后，智能答疑系统会自动将问题及答案加入 FAQ 库，其页面如图 5-10 所示。

主题：什么是队列

提问者：学生一

提问内容：

什么是队列

提交

图 5-6 学生提问问题

主题：什么是队列
提问者：学生一
提问内容：
什么是队列
自动回答：
<p>队列是另一种特殊的表，这种表只在表首（称为队首）进行删除操作，在表尾（称为队尾）进行插入操作。由于队列的修改是按先进先出的原则进行的，所以队列又称为先进先出（First In First Out）表，简称FIFO表。</p>
<input type="button" value="满意"/> <input type="button" value="不满意"/>

图 5-7 系统对问题的自动回答

分治法思想	任何可以用计...	63
什么是树？	树是由一个集...	64
树的表现形式...	父节点数组表...	65
广搜的基本思想	广度优先搜索...	66
什么是队列	队列是另一种...	67
NULL	NULL	NULL

图 5-8 添加到 FAQ 库

主题：什么是队列
提问者：学生一 提问于：2013/2/2 21:36:12
提问内容：
什么是队列
自动回答：
<p>队列是另一种特殊的表，这种表只在表首（称为队首）进行删除操作，在表尾（称为队尾）进行插入操作。由于队列的修改是按先进先出的原则进行的，所以队列又称为先进先出（First In First Out）表，简称FIFO表。</p>
回答

图 5-9 等待教师回答

教师回答:

回答者: 教师一

回答内容:

队列是另一种特殊的表, 这种表只在表首(称为队首)进行删除操作, 在表尾(称为队尾)进行插入操作。由于队列的修改是按先进先出的原则进行的, 所以队列又称为先进先出(First In First Out)表, 简称FIFO表。

图 5-10

图 5-10 教师回答

5.4 系统测试

在完成智能答疑系统的开发后, 对其进行了大量的测试, 其中两项最为关键的测试是中文分词测试和中文句子相似度测试。

5.4.1 中文分词测试

对中文分词进行测试主要是测试中文分词的准确率问题, 在本论文中利用以下公式来计算智能答疑系统中所采用中文分词算法的分词准确率: $\text{准确率} = \frac{\text{正确切分的词汇数}}{\text{总词汇数}}$, 表 5-1 为实验结果。

表 5-1 中文分词测试结果

样本数	切分正确率
100	93
1000	941
10000	9305
100000	92178

另外, 对于输入的包含中英文的自然语言, 图 5-11 为得到的分词结果。

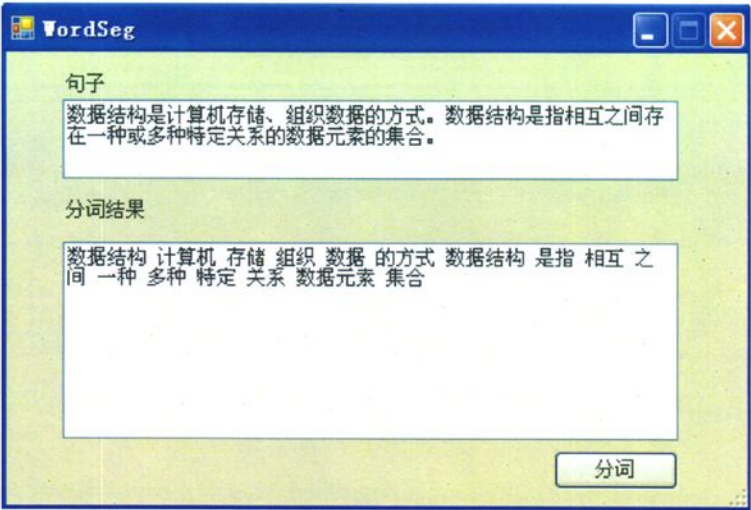


图 5-11 中文分词结果测试

5.4.2 中文句子相似度测试

对中文句子进行相似度测试,主要是利用以下公式来测试中文句子相似度准确率: 准确率=自动答对的题目/问题总数,表 5-2 为实验结果。

表 5-2 中文句子相似度测试结果

问题总数	自动答对的题目
500	412
5000	4600
50000	47025
500000	421788

对于学生用户向提出的问题,系统会在 FAQ 库中寻找相似度最大的问句,并计算两个句子间相似结果和问题所对应的答案。

学生用户可以对自动回答的结果选择满意或者不满意。如果满意,系统就会将此问题和答案加入 FAQ 库。再次提问相同的问题则可直接回答。

若对自动回答的结果不满意或者对于常问问答库中没有找到相关答案的情况下,则需要教师用户来帮忙解答。教师用户可以通过后台查看未解决的问题,并提供解答,其页面如图 5-12 所示。

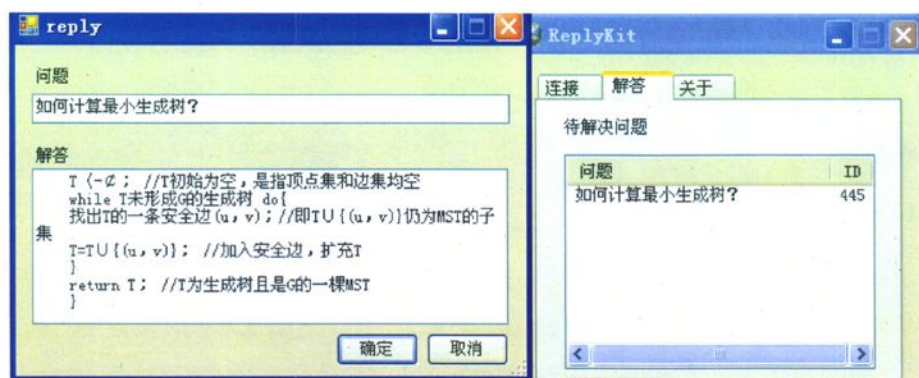


图 5-12 后台管理界面

5.5 本章小结

本章是系统实现与测试部分, 主要对系统的实现过程进行详细介绍, 并对系统进行全面地功能测试等。

第6章 总结与展望

本章对论文的内容进行总结,并对所设计与实现的智能答疑系统存在的不足进行展望。

6.1 总结

互联网技术的发展推动远程教育模式应运而生,通过有机结合互联网技术和计算机辅助教学技术进行教学改革,使得远程教育模式渐成现代教育发展新趋势。智能答疑作为远程教育系统的重要模块,是用来评估远程教育系统优劣性的重要指标,其打破传统的时空限制,进行在线实时答疑。智能答疑综合运用自然语言处理、信息检索等多学科知识,为教师与学生间的交流互动提供了保证。

首先,阅读大量智能答疑系统的相关文献,弄清系统开发的背景、目的和意义,并提出智能答疑系统的解决方案。其次,详细分析系统的用户需求、功能需求和非功能需求,构建完善的体系结构,设计数据库表结构。接着,详细研究在智能答疑模块中如何应用中文信息处理技术,比如自动分词技术、句子语义相似度计算等。最后,为提高FAQ库的查询效率,本文进一步研究数据库快速定位技术,通过比较正排索引结构和倒排索引结构,结合智能答疑系统所涉及的数据特点,最终决定采用倒排索引结构来进行FAQ库的检索,提高关键词检索的效率和精确度。系统测试结果表明,本智能答疑系统已基本满足系统需求。

6.2 展望

对于答疑系统来说,存在的主要问题是系统的智能程度有待提高,答疑系统暂且只能从已有的问答库中搜索相关答案,缺乏思维和推理能力,另外,目前答疑准确率比较低,所以智能答疑系统还有很多可以进一步发展和提升的空间,相信在不久的将来答疑系统能取得重大的突破。

参考文献

- [1] 冯琳, 张爱文. 中国社会发展进程与远程教育的价值取向和功能作用——第五次“中国远程教育专家论坛”综述 [J]. 中国远程教育(综合版), 2009, (6): 5-14.
- [2] 刘照然. 远程教育中智能答疑系统的研究与实现 [D]. 西安: 西安电子科技大学, 2010.
- [3] 郝丹. 中美远程教育研究发展与当前热点的比较研究——以远程教育学术期刊为视角 [J]. 中国远程教育(综合版), 2012, (4): 40-48.
- [4] Don Olcott, Jr. Going Global: Perils and Promises for Open and Distance Education. OPEN EDUCATION RESEARCH, vol. 15, no. 2, pp. 67-71, 2009.
- [5] LH. Yu, SP. Wang. A Research Summary on Quality Improvement in Modern Distance Education. In Proceedings of 2010 Second International Workshop on Education Technology and Computer Science, vol. 3, pp. 735-739, 2010.
- [6] 钟哲辉. 基于计算机网络的信息检索 [M]. 北京: 电子工业出版社, 2007.
- [7] 潘晓辉, 刘志镜. 数据挖掘在智能答疑模型中的应用 [J]. 微电子学与计算机, 2006, 23(3): 116-118.
- [8] 蒋福德, 钟诚. 智能化网络学习系统关键技术研究开发与 [J]. 现代计算机(专业版), 2010, (12): 63-68.
- [9] YJ. Liang, LJ. Zhang, LJ. Ma, and QL. Miao. Research and application of information retrieval techniques in Intelligent Question Answering System. In proceedings of 2011 3rd International Conference on Computer Research and Development, vol. 2, pp. 188-190, 2011.
- [10] SN. Qu, SJ. Wang, Y. Zou and Q. Wang. Research and Design of Intelligent Question Answering System. In proceedings of 2008 International Conference on Computer and Electrical Engineering, pp. 711-714, 2008.
- [11] 阴桂梅, 郭广行. 智能答疑系统模型设计 [J]. 电脑开发与应用, 2011, 24(7): 3-5.
- [12] 孙姜燕. 智能答疑系统中的答案库设计 [J]. 价值工程, 2012, 31(9): 139.
- [13] 程节华. 基于 FAQ 的智能答疑系统中分词模块的设计 [J]. 计算机技术与发展, 2008, 18(7): 181-183.
- [14] <http://www.anyang-window.com.cn/what-is-the-c-s-and-b-s/>
- [15] N. Wang, LM. Li, YZ. Wang, YB. Wang, and J. Wang. Research on the Web Information System Development Platform Based on MVC Design Pattern. In proceedings of 2008.

- International Conference on Web Intelligence and Intelligent agent Technology, vol. 3, pp. 203-206, 2008.
- [16] L. Gao. An Intensive MVC Design Pattern Based on ASP.NET. In proceedings of 2010 Second International Conference on Computer Engineering and Application, vol. 1, pp. 679-682, 2010.
- [17] 许嘉璐, 朱小健. 中文信息处理研究工作的新进展 [J]. 云南师范大学学报(哲学社会科学版), 2010, 42(4): 1-6.
- [18] 黄昌宁, 赵海. 中文分词十年回顾 [J]. 中文信息学报, 2007, 21(3): 8-19.
- [19] 任丽芸. 搜索引擎中文分词技术研究 [D]. 重庆: 重庆理工大学, 2011.
- [20] 兰冲. 基于统计规则的中文分词研究 [D]. 陕西: 西安电子科技大学, 2011.
- [21] 陈桂林, 王永成, 韩客松, 王刚. 一种改进的快速分词算法 [J]. 计算机研究与发展, 2000, 37(4): 418-424.
- [22] 谭琼, 史忠植. 分词中的歧义处理 [J]. 计算机工程与应用, 2002, 38(11): 125-127.
- [23] 刘延吉. 基于词典的中文分词歧义算法研究 [D]. 吉林: 东北师范大学, 2009.
- [24] 翟凤文, 赫枫龄, 左万利. 字典与统计相结合的中文分词方法 [J]. 小型微型计算机系统, 2006, 27(9): 1766-1771.
- [25] 叶继平, 张桂珠. 中文分词词典结构的研究与改进 [J]. 计算机工程与应用, 2012, 48(23): 139-142.
- [26] 张旭. 一个基于词典与统计的中文分词算法 [D]. 四川: 电子科技大学, 2006.
- [27] HS. Wang, and MM. Cui. A Chinese Word Segmentation Based on Machine Learning. In proceedings of 2009 First International Workshop on Education Technology and Computer Science, vol. 2, pp. 610-613, 2009.
- [28] 魏莎莎. 一种中文未登录词识别及词典设计新方法 [D]. 重庆: 西南大学, 2011.
- [29] 郭瞳康. 基于词典的中文分词技术研究 [D]. 黑龙江: 哈尔滨工业大学, 2010.

致谢

大学毕业两年后开始研究生课程的学习，重温学生时代的乐趣。在工作之余兼顾学习，为忙碌的工作生活增添一份乐趣。时光飞快，转眼间我们已经完成所有课程的学习，即将毕业。

在本论文即将完成之际，首先我要向我的导师王美红老师致以最衷心的感谢。本篇论文，从确定课题的研究方向，到选题到开题报告，以及整篇论文的撰写及修改，都是在王老师的悉心指导下完成的。王老师多次询问研究过程，并帮我开拓研究思路、指点迷津，使我受益匪浅。

此外，我还要感谢厦门大学的各位老师尽职尽责为我们授课，让我们在学习中增长了知识，开拓了视野。

衷心感谢我的父母对我的帮助，由于撰写论文期间正好赶上产后不久，如果没有他们帮忙照顾孩子，我将无法在工作之余进行论文写作工作。

另外，我还要感谢所有在学习和生活上帮助过我的同学和朋友，他们的帮助是我不断学习的动力。

最后，衷心感谢在百忙之中评阅本文和参加答辩的各位老师，感谢你们的悉心指导和辛勤劳动。