重庆师范大学

专业硕士学位论文

基于语义相似度计算的智能答疑系统

邢 政

指 导 教 师：魏 延 教授

学 习 形 式：全日制

专业学位类别：工程硕士

专业学位领域：计算机技术

二〇二〇年 五月

重庆师范大学硕士学位论文

基于语义相似度计算的智能答疑系统

硕士研究生：邢 政

指导教师：魏 延 教授

学科专业：计算机技术

所在学院：计算机与信息科学学院

重庆师范大学

二〇二〇年 五月

A Thesis Submitted to Chongqing Normal University in Partial Fulfillment of the Requirements for the Degree of Master

Intelligent Question Answering System Based on Semantic Similarity Calculation

Candidate：Xing Zheng

Supervisor：Wei Yan　Professor

Major：Computer Technology

College：College of Computer and Information Science

Chongqing Normal University

May，2020

基于语义相似度计算的智能答疑系统

摘 要

伴随着互联网技术的不断发展，越来越多的文本信息充斥在我们的生活中。如何在海量信息中快速挖掘出我们所需要的目标信息，成为了我们提高工作效率的重中之重。语义相似度计算（Semantic Similarity Calculation）凭借其对文本间相似程度的准确计算和计算结果的具体量化显示，成为了自然语言处理领域的重要组成部分。该方法在文本分类、信息检索、同义词测试、问答系统等方面起着举足轻重的作用。

目前，在全国大力推行“互联网+”的背景下，“互联网+教育”应运而生。由于网络教育存在地理位置的分割性、师生时间的不一致性，使得教学过程中极为重要的答疑环节成为阻碍网络教育发展的瓶颈。但现有的答疑系统往往仅通过数据库检索或者人工答疑的方法进行答疑，答疑准确率和实效性存在一定的问题。因此，本文对语义相似度相关算法进行了研究，并实现了一个基于语义相似度计算的智能答疑系统。本文具体研究工作如下：

（1）智能答疑算法研究与应用。首先是对本实验中的各个算法进行的详细的对比和介绍，包括中文分词词典的构建、三种常用的基于字符串匹配的中文分词算法、向量空间模型算法。其中核心部分为提出一种改进 TF-IDF 权重计算方法对分词后的各个特征项的权重进行计算，在计算出权重后利用向量空间模型中的余弦相似度的计算方法对问题间的相似度进行计算；

（2）基于语义相似度计算的智能答疑系统的实现。本文在完成算法研究后，实现了一个智能答疑系统。首先对系统进行了分析与设计，包括系统的需求分析、概要设计、功能模块设计、数据库设计等环节；随后完成了系统的实现与测试，系统已基本实现预期的全部功能，包括：自动答疑、教师辅助答疑、学生教师身份互换、相似问题推荐等功能。经过测试，该系统基本可以满足课程答疑的日常需求。

关键字：语义相似度计算，向量空间模型，中文分词算法，智能答疑系统，TF-IDF 权重计算法

Intelligent Question Answering System Based on Semantic Similarity Calculation

ABSTRACT

With the continuous development of Internet technology, more and more text information is flooding our lives. How to quickly dig out the target information we need from the massive information has become our top priority in improving work efficiency. Semantic Similarity Calculation (Semantic Similarity Calculation) has become an important part of the field of natural language processing with its accurate calculation of the similarity between texts and the specific quantitative display of the calculation results. This method plays an important role in text classification, information retrieval, synonym testing, question answering system and so on.

At present, under the background of vigorous promotion of "Internet +" across the country, "Internet+education" came into being. Due to the segmentation of geographical location and the inconsistency between teachers and students, online education makes the extremely important question and answer link in the teaching process a bottleneck that hinders the development of online education. However, existing question answering systems often answer questions only through database search or manual question answering, and there are certain problems in the accuracy and effectiveness of answering questions. Therefore, this paper studies the semantic similarity related algorithms and implements an intelligent question answering system based on semantic similarity calculation. The specific research work of this article is as follows:

(1) Research and application of intelligent answering algorithm. The first is a detailed comparison and introduction of the various algorithms in this experiment, including the construction of Chinese word segmentation dictionary, three commonly used Chinese word segmentation algorithms based on string matching, and vector space model algorithm. The core part is to propose an improved TF-IDF weight calculation method to calculate the weight of each feature item after word segmentation. After calculating the weight, use the cosine similarity calculation method in the vector space model to calculate the similarity between the problems Calculation

(2) Realization of intelligent question answering system based on semantic similarity calculation. After completing the algorithm research, this paper implements

an intelligent question answering system. Firstly, the system was analyzed and designed, including system requirements analysis, summary design, function module design, database design and other links; then the system was implemented and tested, and the system has basically realized all the expected functions, including: Teacher-assisted question answering, student teacher identity exchange, similar question recommendation and other functions. After testing, the system can basically meet the daily needs of course question answering.

Keywords： Semantic similarity calculation, vector space model, Chinese word segmentation algorithm, intelligent question answering system, TF-IDF weight calculation method

目 录

1 绪论

1.1 系统研究的背景和意义

1.1.1 研究背景

随着计算机的普及程度越来越广泛和“互联网+”的普及，各行各业都在借助计算机相关技术来实现自身的发展。而这其中，“互联网+教育”的网络教育模式也成为了目前教育行业较为热门的一种形式。我国网络在线教育行业的市场规模也在近几年取得飞速发展，如图 1.1 所示。网络教育是一种以学生为主体，以计算机、多媒体、Inernet 网络等技术为媒介，运用图像、文字、音频、视频相结合的一种新型教育模式。网络教育相较于传统教育模式，其优势在于突破的地域和时间的限制，学生可以在任何时间、任何地点进行学习，提供了极高的便利性。

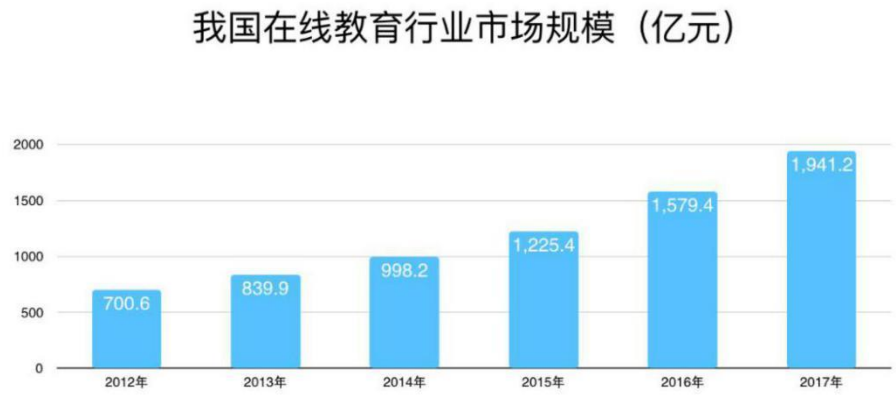

图 1.1 我国在线教育市场规模

其中，作为国内网络教育最具代表性的慕课，从 2013 年进入中国以来，无论是课程数量上还是学习人数上都有了突破性的提升，如图 1.2、图 1.3 所示。与 2017 年相比，短短两年间，网络教育中的各项指标都完成了跨越式发展，我国慕课数量增长近 3 倍、学习人数增长 2.7 倍：慕课数量从 3200 门增至 1.25 万门，学习人数则从 5500 万人次激增至 2 亿多人次；除此之外，国家精品在线开放课程数量也从 17 年的 490 门增加到 19 年的 1291 门，同比增加 1.6 倍（数据来源自 2019 年 4 月 9 日中国慕课大会中国教育部副部长钟登华会上讲话）[1]。由这些数据不难看出，我国网络教育在近两年得到了飞速的发展。

图 1.2 2012 至 2019 年慕课课程数量

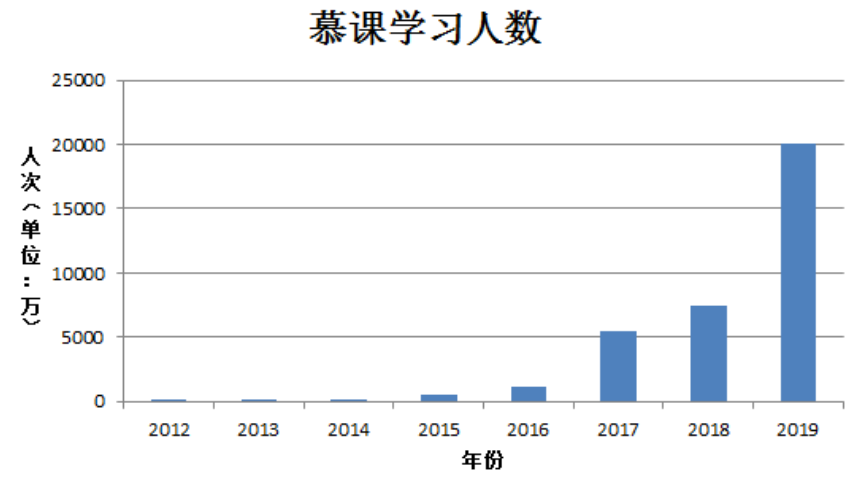

图 1.3 2012 至 2019 年慕课学习人数变化

随着“互联网+教育”模式[2]的不断普及，越来越多的学习者选择以互联网为媒介的自主学习的学习方式，通过互联网的便利性可以进行个性化的学习。无论是传统的教育模式还是现行较为普及的网络教育模式，师生交流始终都是整个教育环节中不可或缺的重要组成部分，而学生提问老师答疑的教育模式也普遍被认为是提高教育教学水平、学生学习成绩的重要方式。在传统的教育模式中，学生出现问题时，可以及时与老师进行沟通交流，老师可以随即为学生进行答疑，整个答疑的过程简单、直接、高效，这种模式是答疑过程中最为理想的方法；而在网络教育中，由于师生之间可能存在时间上的不同步性以及地理位置上的分隔性，可能会出现当学生有问题时老师不能及时进行解答的情况，这种现象无疑会影响到学生接下来的学习进度和学习效率，这就凸显了智能答疑系统的重要性[3]。

智能答疑系统是一个能够远距离的为学生提供本课程内大多数问题的解答，

能够实现老师与同学们之间的互动交流及同学与同学间的相互交流，用来代替传统的教师对某一专一问题的反复解答，在一定程度上减轻了教师的工作强度；同时，该系统也是一个为学生提供高效率、高质量的解决问题的方法和途径。除此之外，智能答疑系统也可以对学生的问题进行总结归纳，把学生学习中存在的问题进行整合，把学习中的重点、难点用数据化的形式加以呈现，可以在一定程度上增强教师的教学质量[4]。智能答疑系统可以看做为网络教育顺利进行提供更进一步的支持，也是教育教学与现代科技相结合的新时代的产物。

1.1.2 研究意义

本文以基于语义相似度计算的智能答疑系统为主要的研究对象。该研究将传统的师生答疑方式与学习者互相交流互相答疑的方式相结合，使得整个答疑过程更加快捷准确，改善了目前答疑系统的不准确性和不及时性的弊端。智能答疑系统的研究意义是消除网络学习中存在的答疑时间上的不及时性和地理位置上的分割性、提高网络教育中答疑的准确性和及时性，是对目前网络教育所存在的问题的一个很好地补充。

通过智能答疑系统的构建，学习者可以利用该系统迅速的解决自己存在的问题，从而可以很好的提高学习的效率。同时，该系统可以在一定程度上为教师提供帮助，可以辅助教师教学，来提高其教学水平和教学质量。

1.2 国内外研究现状

智能答疑系统（Question Answering System，QAS）是一种高效的答疑系统，其立足于传统的自动答疑系统[5]，融入更加智能的答疑方式来为学习者进行答疑，提供更加准确的答案[6]。

1.2.1 国外研究现状

在该领域，因为国外的计算机水平发展较高，及国外语言的语法相对简单，国外在系统上的发展程度比国内更为先进和成熟。国外从上世纪 60 年代就开始有对问答系统的研究，在 1999 年文本检索会议(Text Retrieval Conference，简称 TREC）首次引入问答系统评测专项后，自动问答技术实现了飞速的发展；从 1999 年到 2007 年，这段时间内的自动问答技术的核心工作是通过文本检索，来为用户查找答案；2007 年至今，自动问答技术有了突破式的发展，有人提出了基于网络链接或者元数据结果的问答系统；近年来，随着网络技术的进一步发展，出现了

一种社区问答系统，该系统以常见问答对(Frequent Asked Questions，FAQ)为知识库为用户进行答疑，相较于之前的通过信息检索来进行答疑的问答系统更具有高效性。但国外的答疑系统在功能上相对来说更为精简，基本上所有的功能都是以答疑为核心进行设计的[7]。

Askjeeves 公司的 Askjeeves for kids[8]是于 1996 年设立的，该系统[9]主要是面向 7-14 岁儿童，其答疑方式为目录式搜索引擎。在其主页面上，包含科学、历史、艺术等数十个分类板块，在各版块下又分设更加具体的分类目录。该网站的设计十分贴合儿童的读写习惯，通过逐步与用户进行交互的方式，进行精准的答疑。

MIT 人工智能实验室研发的 START[10] (Syntacti CAnalysis Using Reversible Transformation)，该系统[11]不同于传统的问答系统，其不仅根据用户浏览的频率来提供点击量列表，同时也主要是为了通过系统与用户之间的交互来为用户提供最为准确的信息。该系统的问答范围也十分广泛[12]，包括地理、科学、艺术、环境、历史、文化等多个方面，答疑结果上也是呈现多样化的回答方式，包括：图片、动画、文字等多种形式。

美国 Miehigan 大学的 Answerbus[13]是一个开放领域的问题回答式系统，是基于国外五个主流搜索引擎和主题目录建立的，分别为：Google，Yahoo，WiseNut，Altavista，Yahoo News。同时，该系统[14]支持英语、德语、意大利语、法语、西班牙语、葡萄牙语六种语言，其通过在多个搜索引擎进行答案检索后，为用户进行答疑，实现了基于句子层面的信息检索。

通过对国外各系统的研究[15]总结，发现国外的智能答疑系统主要有以下特点：首先，国外研发的系统普遍具有很强的独立性，对不同形式的答疑问题进行回答时，经常将其作为独立的模块进行检索，系统独立性极强；此外，国外答疑系统基于语言语法相对简单的特性，其智能性体现在当用户提交较为准确的问题时能够给出准确的回答，对用户问题的归纳有很高的要求，一旦出现问题问法不准确时答疑效果会大打折扣。

1.2.2 国内研究现状

在智能答疑系统这一领域，国内研究起步相对较晚，但是近年来，发展势头迅猛。1998 年，上海交大的周睿斌等人提出了 Answer Web 自动答疑系统[15]，是我国第一次提出自动答疑的提法，该系统是将学生的提问根据课本的章节内容进行解答，这样不利于学生对于所学全部知识的融合；江西科技师范大学的丰乃波提出了一种以学习者为中心的智能答疑系统[16]，该系统能够较好的处理学生的问题，但是没有对问题与问题之间的语义建立联系；除此之外，北京师范大学的 Vclass

平台[17]，通过自身完备的知识库，建立了一个能够为学生精准答疑的平台，但是也没有对通过问题间的相似度对其进行相似问题推荐。经研究发现，现行较为流行的智能问答系统主要采用简单的数据库查询或人工答疑，或者是将某一门学校课程与答疑系统进行简单的结合后，定向的针对学生提出的问题由数据库进行检索或教师答疑，但会存在准确度不高或者人工答疑时等待时间较长的问题，故建立智能答疑系统，是在未来一段时间内答疑系统发展的主流趋势。

然而，由于汉语和英语在语法、使用习惯等多方面存在较大的差异，国外的答疑系统无法很好的匹配汉语语言的使用，这就导致了目前国内的答疑系统并不完善；另一方面，国内大多从技术的角度开展研究，仅仅依靠几个关键词进行系统检索，并非能智能的对用户提出所有的问题都能解决，这就要求了用户在提问时要自行提炼问题中的关键字，而并非所有提问的用户都能完成此工作，且目前得智能性的答疑系统中，系统的专业性和智能性还有待提高。

1.3 研究内容

答疑过程对于提高学生的学习效果、改善教师的教学方法有着重要意义，针对目前网络教育过程中存在答疑不准确、不及时的现象，本文研究并实现了基于语义相似度计算的智能答疑系统。本文主要研究内容包括：

（1）对中文分词技术中的分词方法进行研究比对，选出其中最适用于答疑系统的逆向最大匹配分词方法，对其词典构造、分词过程进行详细研究，并在系统中利用该方法对学生输入的问题进行分词处理，将长问题处理成为较短的关键词作为特征项进行答案匹配；

（2）对语义相似度算法中的向量空间模型算法进行研究，重点研究各个特征项输入后的权重计算方面。在本研究中，提出了一种加入权重因子的改进 TF-IDF 权重计算法，提高了各个特征项权重计算的合理性，并将该算法运用到系统中，利用该计算方法来计算学生输入问题与数据库中各个问题间的相似度，为学生完成精准答疑；

（3）在本研究中，将《数据结构》课程的问题作为实验数据来进行研究。在实验中基于逆向最大匹配算法和向量空间模型算法，实现一个可以准确答疑的智能答疑系统。

1.4 论文组织结构

本文将分为六个章节进行论述，各章节的具体安排如下：

第 1 章为绪论，该部分主要论述该系统的研究背景及意义、智能答疑系统的

国内外研究现状，同时对整个文章的研究内容进行了综述，对文章的组织结构进行了说明；

第 2 章为语义相似度相关理论基础，主要讲述的是语义相似度计算的总体概述，主要描述了目前常用的三种语义相似度计算方法：向量空间模型、布尔模型和概率模型；随后将三种常用的分词方法进行了对比，并重点介绍了本文使用的中文分词技术中的基于字符串匹配的中文分词算法；同时也对常用的四种权重计算法进行了分析说明；

第 3 章为智能答疑算法研究与应用，主要讲述的是整个系统进行自动答疑的主要流程及算法使用：首先说明了基于字符串匹配的中文分词算法，着重介绍了构建分词词典的方法和逆向最大匹配算法的使用；接着对本文语义相似度计算的流程进行了阐述，分别为文本向量化、特征项权重计算、问题相似度计算及答案提取，其中在特征项权重计算中提出了本系统中使用的改进 TF-IDF 特征项权重计算方法；

第 4 章为基于语义相似度计算的智能答疑系统的分析与设计，包括系统的学生用户需求分析、教师用户需求分析、可行性分析、系统的概要设计、用户登录模块设计、学生用户模块设计、教师用户模块设计及数据库设计；

第 5 章为基于语义相似度计算的智能答疑系统的实现与测试，主要阐述了系统的开发环境，以及系统基本功能的实现和系统的测试结果的展现；

第 6 章为对本文工作的总结与展望，对全文进行的工作进行了总结，并指出了在实验过程中的不足之处以及下一步继续研究的重点和方向。

2 语义相似度相关理论基础

2.1 语义相似度计算

语义相似度计算是自然语言处理领域重要的基础研究之一[18]。在目前信息爆炸的时代背景下，我们接受到的文本信息数不胜数。对于这些非结构性的信息，利用文本挖掘的相关技术对其进行分析整理可以在很大程度上提高我们的工作效率，而语义相似度计算正是该工作最重要的方法之一。语义相似度计算[19]是指将计算机技术与语言学的相关知识进行融合后，通过一定的技术手段，对两个及以上的实体（包括语句、短语、文档等）之间的相似程度进行计算，并得到一个具体数值的过程[20]。语义相似度计算即通过对比多个文本之间的相似性后，来实现信息检索、文本分类、摘要提取等功能，除此之外，语义相似度计算在自然语言理解与处理、知识获取、同义词探测、信息抽取、机器翻译等多个领域都有着广泛的应用。

就目前的发展状况来说，语义相似度计算方面还未形成对于相似度统一的度量方法，一般来说都是用[0,1]之间的一个数来衡量。在国外对语义相似度计算的使用中，常用的方法即为通过对文档中最小的单元进行检索对比后进行计算，而在中文文本中，最小的单位为单个的汉字，显而易见的是，对单个汉字的研究并没有多大的价值和意义[21]。因此在对中文文本的研究中，通常是以词语、句子、段落为基本单位来计算文本中不同的词语、句子、段落的相似程度，并对其进行相似度计算后，通过具体的数值对文本间的相似性进行评估的过程。在目前这样的发展背景下，语义相似度计算在多个方面都起到了重要的作用，而其在智能答疑系统这一方面更是有着不可代替的作用。

目前来说，语义相似度计算有多种方法，主要包括：向量空间模型、布尔模型、概率模型三种模型算法。

2.1.1 向量空间模型

向量空间模型（Vector Space Model，VSM）是一种常用的文本语义相似度比较算法[22]，是 Gerard Salton 等人在 1969 年提出的，该模型也是目前相似度计算中应用最为广泛的模型[23]。该模型在信息检索、文本信息分类、文本聚类等方面都有着很重要的作用，在自然语言处理领域[24]也有着不可替代的作用。

在该模型中，使用该模型的前提是假设某特征在文本中的作用，仅与其在文本中的频数有关，与其所在文本中的顺序、位置无关。在对文本进行处理时，按

照其文本中相似的特征值的频数对其进行处理，不考虑其在文本中的位置。在对其进行相似度计算时，也是以两个文本之间特征项的相似程度来计算两者间的相似度。

有了上述对向量空间模型的简介，再对该模型的核心思想进行概述：在该模型中，往往将需要进行相似度计算的文档 T_1，T_2，……，T_n 看成是在空间中需要进行计算的向量，如果要对这文档进行相关的相似性计算，只需要对文档对应向量的内积进行计算，如果文档之间的相关性越强，则向量之间的内积越小；如果文档之间的相关性越弱，则向量间的内积越大。

下面首先对向量空间模型中使用到的相关名词和定义进行说明：

①文本：泛指输入的各种文字，可以是语句、词语、段落等形式，是我们日常生活中使用的语言，常用 T_i 来表示；

②特征项：是指能够反映文本中用来表达主题的关键词语，在模型计算中常用 t_i 来表示；

③权重：指的是特征项在本文本中的重要程度，特征项的权重大小往往与其在文本中的重要程度成正比，即越重要的特征项其权重越大。在模型计算中常用 ω_i 来表示。

基于对模型的概念、核心思想、专属名词的理解，我们可根据这些理论，来开展建模工作。建模的基本思想为：在一个 n 维的空间中建立一个空间坐标系，文本中的每一个特征值均代表坐标系中一个维度，分别用 T_i，T_k，T_j 等表示，且各个维度均是相互独立的且各个维度间均存在一个夹角 θ，这个夹角在不同维度间的大小不同，这是进行相似度计算时的关键所在。该空间模型示意图如图 2.1 所示。

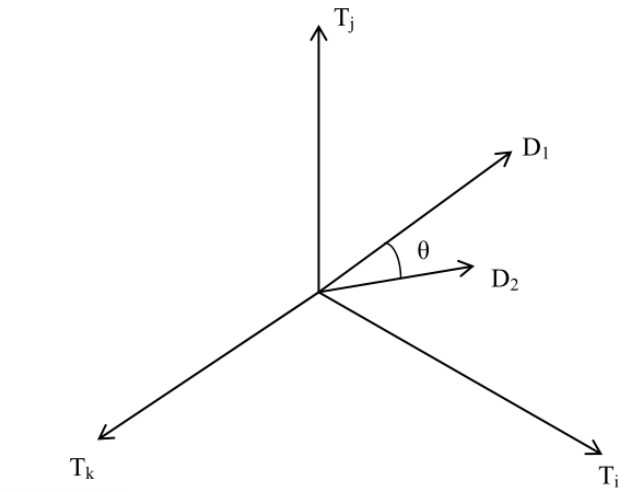

图 2.1 向量空间模型（VSM）示意图

2.1.2 布尔模型

布尔模型（Boolean Model）[25]是由 George Boole 提出的一种基于集合理论和布尔代数进行信息检索和相关度计算的数学模型，它的特点是查找那些于某个查询词返回为“真”的文档。

在该模型中，简单的其默认为一篇文档只有相关和不相关两个状态，在其数学表达式中，特征项值用 0 和 1 来表示，“1”代表相关，“0”代表不相关，特征文本用这些特征项组成的特征向量来表示。布尔模型的数学表达式为：

$$\omega_i = \begin{cases} 1, TF(T_i) \geq 1 \\ 0, TF(T_i) = 0 \end{cases} \quad (2.1)$$

布尔模型的定义非常直观，因此该模型也具有简单、明了、检索速度快等优点，该模型能够用来表达一些结构性信息。但同时，该模型的缺点也是非常明显的：基于二元判定标准设计的过于单一的模型，使模型缺乏文档分级的思想，对文本仅设相关不相关两种关系，明显是不能涵盖所有文本信息的，这极大地限制了布尔模型的使用。总之，该模型在文档检索方面还需要进一步优化使用。

2.1.3 概率模型

概率模型（Statistical Model）[26]是一种基于概率排序原则的相关性评价模型。该模型是指通过计算词与词之间、词与文本之间的概率的文本来表示的模型。该模型根据贝叶斯决策理论，通过概率论中的文本分布来表示文本间的相关性。概率模型具有较高的准确性，但是该模型运算较为复杂，且要求的独立性条件较高，使用总体效果较向量空间模型有一定的差距。

2.2 中文自动分词技术

为了使系统能够更加准确的理解学习者提出的各个问题，利用中文自动分词技术[27]对学习者提出的问题进行处理是系统答疑的第一步。中文自动分词技术就是对语句中的字词进行准确的划分，分词不准确将会导致系统对输入语句的理解错误从而降低答疑的准确性，出现答非所问的情况。目前中文自动分词技术在搜索引擎、问疑系统、推荐系统等各个方面被广泛应用，并取得了很好的进展。现阶段，我国最常用的中文自动分词技术[28]有三种，分别为：基于字符串匹配的自动分词、基于统计的自动分词及基于语义理解的自动分词。

2.2.1 基于字符串匹配的自动分词

基于字符串匹配的自动分词技术[29]又被称为机械分词方法，是目前中文算法

中产生最早的算法。一般该方法在使用时要事先建立一个足够大的分词词典，随后将需要进行分词的语句输入后，与分词词典进行逐条匹配。如果在该字符串存在于词典中，那么就说明匹配成功，并将匹配成功的字符串从原文本中切分出来。该方法是目前自动分词方法中发展最早、最为成熟的一种，也正是由于其算法相对简单、上手难度较低、容易理解等特点，也是目前使用频率最高的方法。但是从算法逻辑的角度不难发现，该算法的准确性及效率都是需要依靠字典的，一般来说，系统出于分词的效率来考虑，并不会使用太大的字典，这样就会出现在字典中的词语并不全面，在匹配过程中会有无法匹配的词语出现，我们将这些词成为未收录的词。未收录词主要分为两类，一类是新的普通词汇或专业词汇，另一类是专有名词，包括中外人名、地名、机构名、事件名等。

基于字符串匹配的中文算法其算法思路为：是按照一定的方式将待分析的字符串与词典中的词条进行配，若在词典中找到某个字符串，则完成匹配并切分该词。目前常用的基于字符串匹配的算法有以下几种：正向最大匹配法（FMM）[30]、逆向最大匹配法（BMM）[31]、双向最大匹配法[32]等。

2.2.2 基于统计的自动分词

基于统计的自动分词方法[33]是随着大规模语料库的建立，统计机器学习方法的发展背景下逐渐建立起来的。该算法的主要思想是将词理解成为是由字组成的，当相连的字在文本中多次出现时，则默认为这样的相连的字就大概率是词。在这种思想下，当某两个相连的字多次出现时，其频率也相应较高，在这个频率高于某一临界值时，就默认该相连的字构成的是词语。

相比于基于字符串匹配的自动分词方法，统计分词方法不需要耗费人力维护词典，能较好地处理歧义和未登录词，但是，其分词的效果很依赖训练语料的质量，且计算量很大，会造成在使用时需要占用大量的系统资源，造成大量的资源开销浪费。

2.2.3 基于语义理解的自动分词

基于语义理解的自动分词技术是将计算机模拟成人对句子进行理解的一种分词方法[34]，其核心思想为在进行分词的同时，对待分词的句子进行语法、句意的分析，从而进一步达到更精准分词的效果。目前该方法下最常用的方法为联想-回溯法，该方法要求建立知识库、特征词库、实体词库和规则库，在进行分词时首先将输入的语句划分成为若干个子串，子串可以是词也可以是句子，然后在利用实体词库和规则库将子串再细分为词，在切分的过程中要充分利用多种汉语语法

知识，才能切分的更加精准。该种方法增加了算法的时空复杂度，且从理论上来说具有更好的效果，但是由于汉语的复杂性和多样性，目前尚处于研究阶段。

2.2.4 三种分词方法对比

在对上述三种自动分词方法进行全方面的对比实验后发现，实验结果如表 2.1 所示。结合本实验的特点，在实验过程中是针对某门课程开设的答疑系统，因此在分词词典构建完善的前提下，并不会存在太多的未登录词需要识别，可以避免基于字符串匹配的分词技术的最大弊端。因此综合衡量各方面因素，在实验过程中拟使用基于字符串匹配的自动分词技术来进行搭建。

表 2.1 三种分词方法对比

分词方法	基于字符串匹配	基于统计	基于语义理解
是否需要词典	是	否	否
是否需要语料库	否	是	否
是否需要规则库	否	否	是
算法复杂性	简单	一般	复杂
技术成熟度	成熟	成熟	研究阶段
分词速度	快	一般	慢
新词识别	差	强	强
实施难度	容易	一般	难度较大
分词准确性	一般	准确	准确

2.3 权重计算法

在对相似度的计算过程中，特征项的权重是来衡量特征项在文本中的重要程度的关键指标[35]。只有当相对重要的词汇占据了相对较大的权重时，对语句进行的相似度计算结果才会具有较高的可信度。目前来说，常用的权重计算方法有 TF 权重计算法[36]、DF 权重计算法[37]、TF-IDF 权重计算法[38]、熵权重计算法[39]等。

2.3.1 TF 权重计算法

TF 权重计算法（Term Frequence）是简单的按特征项在文本中出现的频率作为计算依据来对特征项进行权重计算的。在该算法下，特征项的频率是最重要的因素，即特征项在文本中出现的频率越高，其所占权重越大。这种方法其优势在于计算简单，但是仅适用于计算文本权值时使用，适用范围较小。其计算公式如下所示：

$$W_{tf} = C_t \tag{2.2}$$

在该表达式中，ω_{tf}即为特征项权重，C_t为特征项在文本中出现的次数。

2.3.2 DF 权重计算法

DF 权重计算法（Document Frequence）是通过特征项在文档中出现的频率来区分不同文本时使用的一种算法。当文本规模较大时，用特征项在所有文档中出现的频率来代表特征项的权重，也就是说，特征项的权重直接与其在文档中出现的频率相关，频率越高则其权重越大。其计算公式如下所示：

$$W_{df} = D_t \tag{2.3}$$

在该表达式中，ω_{tf}即为特征项权重，D_t为特征项 t 的文档频率

2.3.3 TF-IDF 权重计算法

TF-IDF 权重计算法（Term Frequence Inverse Document Frequency）是目前使用较为广泛的权重计算法，同时考虑到了特征项在文本中的出现频率（TF）及特征项在文档中的逆文件频率（IDF），该算法是一种基于统计学的方法，可以利用该算法可以来评定选取的字词对于整个文本的重要性。

在该方法中，TF（Term Frequence，词频）指的是某个词在文本中出现的次数。可以表示为字或者词 t_i 在文本 T 中出现的次数，若 t_i 在 T 中出现 k 次，则可记为：

$$TF_{i,k} = n_{i,k} \tag{2.4}$$

IDF（Inverse Document Frequency，逆文件频率）是来判断一个词语对整个文档集重要程度的主要判断依据[40]。它指的是特征项文本 T 中出现的次数越多，则其区分度越低，应该降低其权值。数学表达式可记为：

$$IDF_{i,k} = \log \frac{|D|}{\left|\left\{k : t_{i,k} \hat{I} D\right\}\right|} \tag{2.5}$$

其中|D|是文档集中的所有文档总数，分母表示包含词语 t_i 的所有文档数。假设文档集中有 N 篇文档，特征项 $t_{i,k}$ 在第 $n_{i,k}$ 篇文本中出现过，则

$$IDF_{i,k} = \log \frac{N}{N_{i,k} + \alpha} \tag{2.6}$$

式中 α 为经验常数，取为 0.01，是为了防止当文档中不包含该特征项 $t_{i,k}$ 时分母为 0 的情况出现。

在对 TF-IDF 权重计算方法进行计算时，分别计算出特征项的 TF 和 IDF 值后，将二者结果相乘后作为其权重进行计算，即 TF-IDF=TF×IDF。通过公式可以看出，该算法中，特征项的权重与该其出现频率成正比，与其在整个数据库中出现的次数成反比。

利用此方法，对问题中每个词语的 TF 的 IDF 进行计算后，算出每个词语的权重 w 并保存至数据库中。当有新问题输入时，数据库可以对每个词语的权重进行更新，来保证每次计算时权重都为系统中最新的权重值。

2.3.4 熵权重

熵权重计算法是用来判断某些指标的离散程度的方法。该方法主要是用来衡量信息的离散程度，或者是说不确定度。当某个变量的不确定度越大时，其熵权重也会越大。其计算公式如下所示：

$$H = -\sum_{i=1}^{n} p_i \log p_i \quad (2.7)$$

在该表达式中，n 表示共有 n 种情况，p_i 表示变量 p 为第 i 种情况的可能性。

2.4 本章小结

本章主要对现有的语义相似度计算方法、中文分词算法及权重计算法进行了学习和研究，其中重点研究了本系统中主要使用的语义相似度计算中的向量空间模型算法。在本章的第一部分，首先对现在的语义相似度计算的总体情况进行了概述，分别对现有的三种语义相似度计算方法进行了描述，并重点阐述了向量空间模型算法；在本章的第二部分，对现在比较流行的几种中文分词技术进行了对比分析，随后选出了基于字符串匹配的自动分词技术来使用在本系统中；第三部分则是对四种常用的权重计算方法进行了对比，分析了各种权重计算法的优劣。

3 智能答疑算法研究与应用

经过第 2 章的语义相似度相关理论基础的概述，该系统将以向量空间模型作为对于输入问题处理的主要方法，同时结合基于字符串匹配的中文分词技术对系统进行部署。算法在本系统中的主要应用如图 3.1 所示。

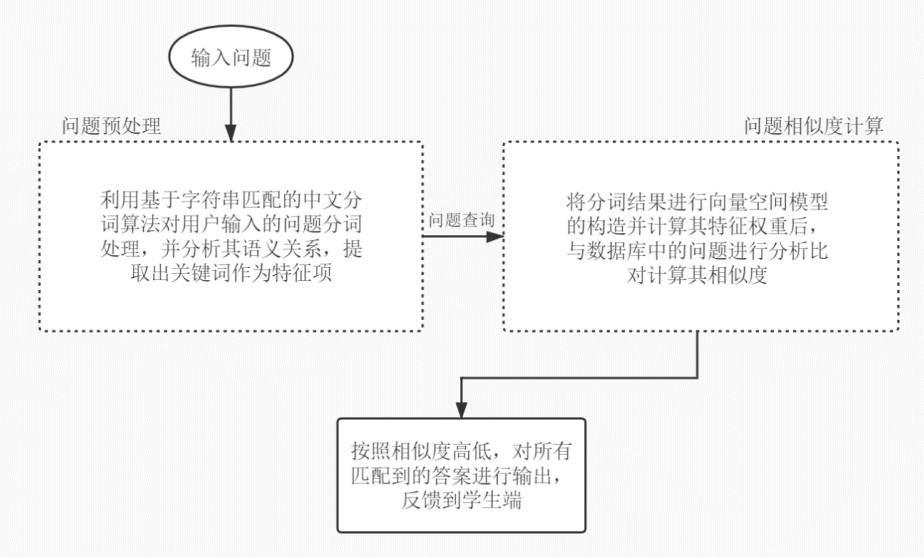

图 3.1 算法应用示意图

3.1 基于字符串匹配的中文分词算法

在答疑系统中，若想保证答疑的准确性，则中文分词算法是保证其准确的前提和基础[41]。在上文中对三种分词方法进行对比后不难发现，基于字符串匹配的算法具有简单、易操作的特点。但是该算法使用的前提是需要构建一个适用于本系统的分词词典，才能提高系统的分词效率和分词的精准度。本节将从构建分词词典和算法的对比及使用两个方面对该算法在本系统中的应用进行阐述。

3.1.1 构建分词词典

在构建分词词典[42]时，分词词典的机制将直接影响系统分词的准确性和高效性，因此必须构建一个分词速度快、准确性高的词典，来提高系统答疑的准确性。目前，常用的构建词典[43]方法如下：

（1）基于整词二分的词典机制

整词二分的词典机制是一种传统的构造分词词典的词典机制。在该机制下，词典由首字索引表、词索引表、词典正文三部分构成的，是利用首字来索引首字在文中的位置，随后根据从该词开始展开，进行二分查找的一个过程。而词典正

文则是根据汉字首字国标码来确定的。在查找匹配的过程中，该机制则是根据对首字进行哈希运算来与词典进行匹配，在词典中查找到该词位置的过程。该机制的结构如图 3.2 所示。

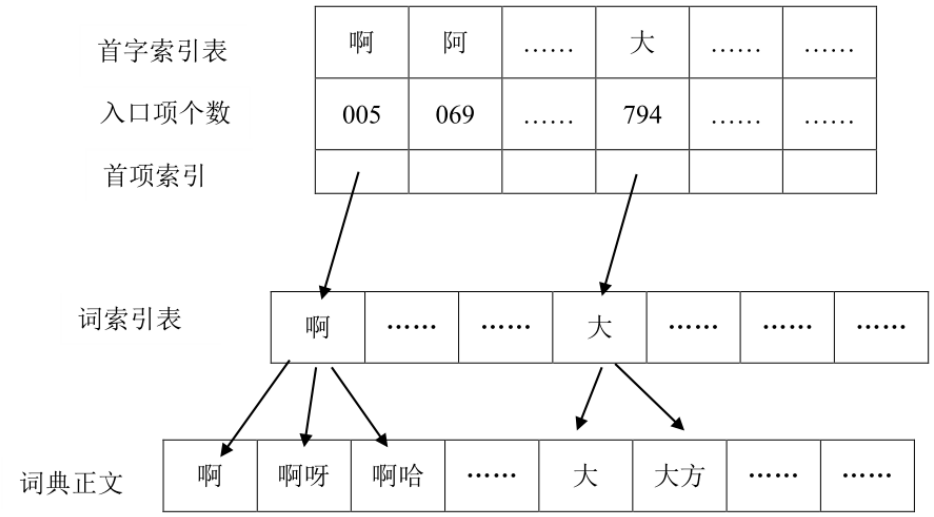

图 3.2 基于整词二分词典结构

以查找 Str=“小李是一个大方的人”为例，从“大”开始进行匹配，根据图 3.2 的基于整词二分词的词典结构，其匹配过程如下：

①取“大”字开头的最长字符串，记为 SMAX，则在本实例中，SMAX= “大方的人”；

②采用整词二分法，在整个词典中对 SMAX 进行匹配，发现未匹配到该词；

③删除 SMAX 中的最后一个字，本例中为“人”，则 SMAX= “大方的”；

④重复步骤②，发现未匹配成功，则重复步骤③，SMAX= “大方”；

⑤在词典中发现“大方”一词，匹配成功，返回 SMAX= “大方”。

在整个匹配过程中，先匹配文中从首字开始的最大的词，以此进行匹配，匹配不成功，删除最后一个字后重复匹配，直至匹配成功。我们通过该过程可以看出，该方法利用了哈希运算的方法在词典中进行快速匹配，避免了重复搜索的过程，因此在匹配速度上具有较大的优势。

（2）基于 Trie 树查询的词典机制

基于 Trie 树查询的词典机制是结合了整词二分法的机制以及 Trie 树的相关理论知识提出的一种词典机制。该机制由两部分组成：首字索引表和 Trie 索引树结点。该机制是在首字的基础上，沿首字逐字进行搜索的过程。该机制的结构图如

图 3.3 所示。

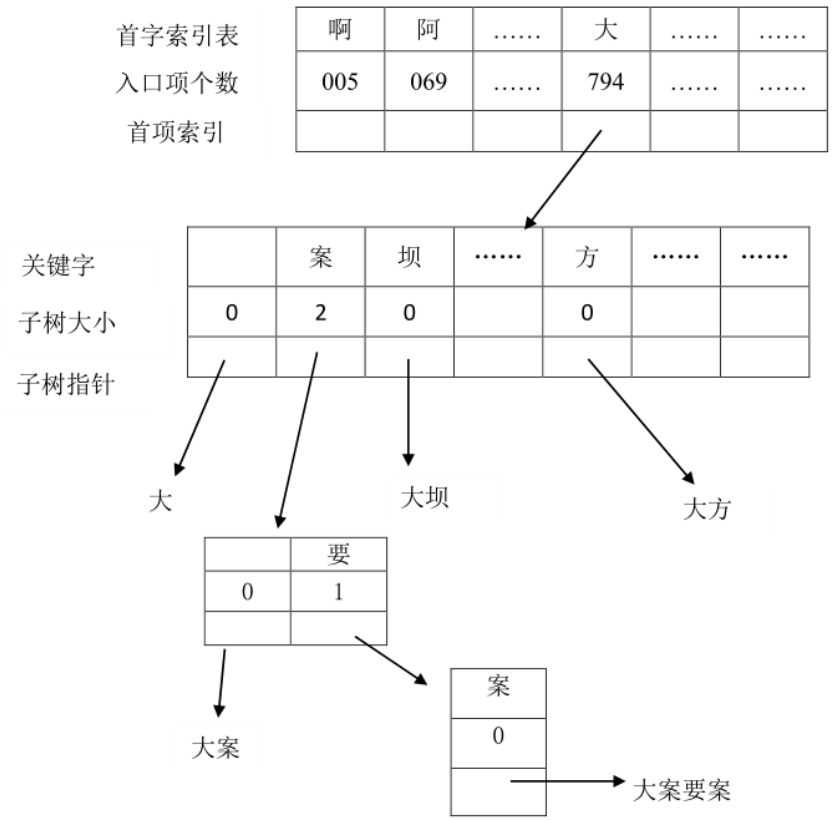

图 3.3 基于 Trie 树查询的词典结构

由结构图不难看出，该机制由于在匹配前不清楚被切分字段的长度，因此在匹配过程中需要从字的根节点开始逐层遍历，这样的查询方法无疑需要大量的空间才能够完成。

（3）基于逐字二分的词典机制

基于逐字二分的词典机制是一种综合整词二分词典机制和 Trie 树查询词典机制的分词机制。在该机制中，其词典结构是与整词二分法的词典结构一致，但是在查询过程中是吸收了 Trie 树查询词典机制中的查询方法，以字为查询单位进行查询，不需要知道词的长度，在查询过程中对输入字符串进行扫描后，对所有出现的词进行匹配查询。以检索“大小年”为例，展示其词典结构如图 3.4 所示。

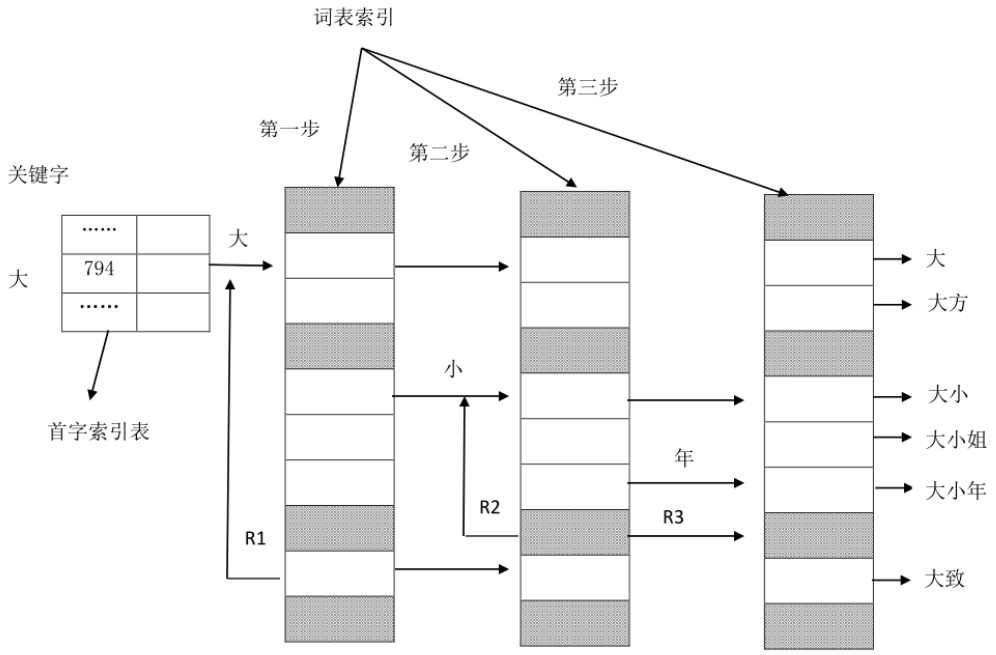

图 3.4 基于逐字二分的词典机制结构

在图 3.4 中可以看出，R_1，R_2，R_3 位分别按上一步检索出的关键字进行再次检索的过程，其检索出的结果之间的关系可以表示为 $R_3 \in R_2 \in R_1$。在基于逐字二分的词典机制中，将基于 Trie 树查询的词典结构中的查询优势运用到基于整词额二分法的词典结构中，整个方法的运行效率有了很大的提升。

（4）分词词典设计

在完成对三种词典结构和特性的分析后，结合本实验中是以《数据结构》课程为例设计的智能答疑系统的特点，在本实验中，采用的是基于逐字二分法的词典构造机制来构建词典。

本系统中，将会把词典的设计分为：专业词典和停用词典两部分。对于本系统来说，专业词典是保证答疑准确率和效率的重中之重。在专业词典中，存放的是《数据结构》课程内的专业词汇，如：拓扑排序、有向无环图、哈夫曼树等。结合学生进行答疑时的输入特点，有可能会出现并非完全的汉语输入方式及部分语气助词，因此通过建立停用词典的方式，删除部分无实际意义的虚词，在检索时不对其进行检索，来提高检索的效率。

3.1.2 分词算法

在本系统中，采用基于字符串匹配的中文分词算法主要是用来对输入的问题进行分词处理，保证系统能够快速对问题进行切分，这是保证后续进行语义相似

度计算的前提。

在上文中已经谈到，基于字符串匹配的中文分词算法有正向最大匹配法、逆向最大匹配法、双向最大匹配法等几种方法。在研究中首先对三种基于字符串匹配的中文分词算法进行对比实验后，选择逆向最大匹配算法进行使用。

（1）正向最大匹配中文算法

正向匹配中文算法是将输入文本从左至右进行分词，在分词过程中将文本中切分词汇与词表进行匹配，若匹配成功，则切分出该词。这种方法在思想上较容易理解，但是存在一定的问题：若匹配过程中，最大匹配到的词必须保证下一个扫描的对象不是词表中的词或者词表中某些词的前缀才能结束。

如采取正向匹配中文算法，对字符串 S=“我们在野生动物园”进行分词，将最大词长设为 6，则分词过程如表 3.1 所示。

表 3.1 正向最大匹配中文算法

步骤	字符串 S	操作	分词结果
1	我们在野生动物园	选取 6 个字符	无
2	我们在野生动	无匹配结果	无
		去掉一个字符	
3	我们在野生	无匹配结果	无
		去掉一个字符	
4	我们在野	无匹配结果	无
		去掉一个字符	
5	我们在	无匹配结果	无
		去掉一个字符	
6	我们	与词典匹配成功	我们
		取出该词	
7	在野生动物园	无匹配结果	我们
		去掉一个字符	
8	在野生动物	无匹配结果	我们
		去掉一个字符	
9	在野生动	无匹配结果	我们
		去掉一个字符	
10	在野生	无匹配结果	我们/在野
		去掉一个字符	

（续表 3.2）

步骤	字符串 S	操作	分词结果
11	在野	无匹配结果 去掉一个字符	我们/在野
12	生动物园	无匹配结果 去掉一个字符	我们/在野/
13	生动物	无匹配结果 去掉一个字符	我们/在野/
14	生动	与词典匹配成功 取出该词	我们/在野/生动
15	物园	无匹配结果 去掉一个字符	我们/在野/生动
16	物	与词典匹配成功 取出该词	我们/在野/生动/物
17	园	与词典匹配成功 取出该词	我们/在野/生动/物/ 园

（2）逆向最大匹配中文算法

逆向匹配中文算法该算法是将输入文本从右至左进行分词，在分词过程中的思路大致与正向最大匹配法一致。采用逆序的方法对文本进行分词，充分利用了汉语的使用习惯，很好的避免了正向最大匹配法存在的问题。

逆向最大匹配分词算法的流程如下所示：

①用户输入后，系统获得一个句子

②设句子中的字数为 n；

③设置我们要截取词的长度，记为 m；

④从句子中取 n-m 到 n 的字符串 subword，去字典中查找是否有这个词。

如果有就进行⑤, 没有就进行⑥；

⑤记住 subword，从 n-m 付值给 n，继续执行④，直到 n=0。

⑥将 m−1，再执行④。

在该过程中，简单来说就是选取词最大长度 max，通过对输入语句从右向左进行遍历的方式，来对句子进行分词。其流程图如图 3.5 所示。

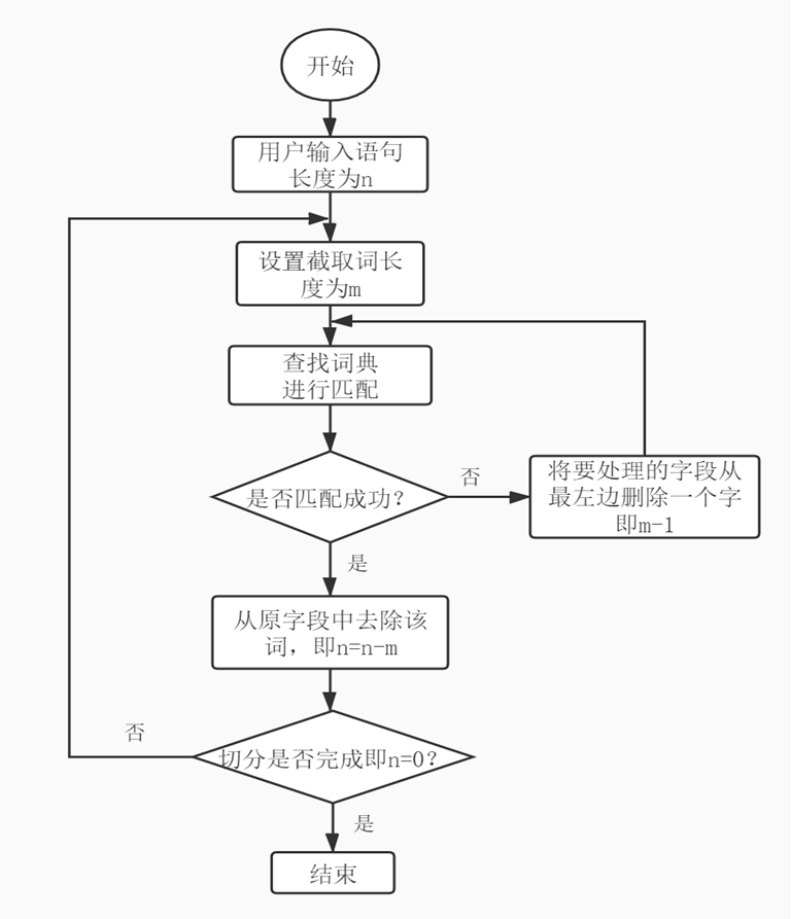

图 3.5 逆向最大匹配分词法流程图

如采取逆向匹配中文算法，对字符串 S=“我们在野生动物园”进行分词，将最大词长设为 6，则分词过程如表 3.2 所示。

表 3.2 逆向最大匹配中文算法

步骤	字符串 S	操作	分词结果
1	我们在野生动物园	选取 6 个字符	无
2	在野生动物园	无匹配结果	无
		去掉一个字符	
3	野生动物园	与词典匹配成功	野生动物园
		取出该词	
4	我们在	无匹配结果	野生动物园
		去掉一个字符	
5	们在	无匹配结果	野生动物园
		去掉一个字符	

（续表 3.3）

步骤	字符串 S	操作	分词结果
6	在	与词典匹配成功 取出该词	在/野生动物园
7	我们	与词典匹配成功 取出该词	我们/在/野生动物园

（3）双向匹配最大分词算法

该算法是综合正向匹配算法和逆向匹配算法提出的一种算法，在分词过程中，使用一次正向匹配得出结果后，再使用一次逆向匹配，若二者结果相同，则直接使用；若结果不同，则看分词结果数量来确定：若分词数量不同，则采用分词数量较少的作为分词结果；若分词数量相同，则采用单字数较少的作为分词结果。该算法虽然提高了分词的准确度，但是实体词长度大于窗口大小时会影响分词结果，对分词长度有较高的要求。

（4）三种分词方法对比

由上述分词过程可以看出，双向匹配最大分词法在空间上要求更高，且对分词长度的要求较高，因此在实验中不采用此方法；在对正向匹配中文分词算法和逆向匹配中文分词算法的对比中发现，同样输入字符串 S=“我们在野生动物园”时，逆向匹配中文算法的分词效率要明显高于正向匹配分词算法，同时，由于汉语词语的特性，逆向匹配最大分词算法的结果往往更贴合于我们想要表达的含义。同时根据大量实验显示[44]，正向匹配中文分词算法的错误率约为 1/169，而逆向匹配一般错误率约为 1/245。因此在本研究中，选用的是逆向最大匹配中文分词算法。

3.2 语义相似度的计算

在本章的第一节中，我们已经完成了对输入问题的分词工作。在此节内容中，我们通过对输入问题的权重计算、语义相似度计算、答案提取等工作，来完成对问题的语义相似度计算。

3.2.1 文本向量化

完成首先要进行的工作是对输入文本的向量化。假定输入文本 T，则在该文本内确定存在 n 个不同的特征项$\{t_1,t_2,\ldots\ldots,t_n\}$，这些特征项在文本中的表示为：

$$T\{t_1,t_2,\ldots\ldots,t_n\} \quad (3.1)$$

在公式 3.1 中，每个特征项 t_i 在整个文本 T 中都对应其相应的权重 ω_i，放到 n 维坐标系中来说，特征项 t_i 对应的是坐标系的坐标轴，而权重 ω_i 代表的则是各个

坐标轴上的坐标值，这样，就可以将文本 T 在空间坐标系中进行表示。

$$\overrightarrow{V_T} = (\omega_1, \omega_2, \cdots\cdots, \omega_n) \tag{3.2}$$

式中的 $\overrightarrow{V_T}$ 代表的是文本 T 的特征向量。

结合式 3.1 和式 3.2，可以得出文本 T 的向量矩阵为：

$$T = \begin{bmatrix} \overrightarrow{V_1} \\ \overrightarrow{V_2} \\ \vdots \\ \overrightarrow{V_n} \end{bmatrix} = (t_1, t_2, \cdots\cdots, t_n) = \begin{pmatrix} \omega_{11} & \cdots & \omega_{1n} \\ \vdots & \ddots & \vdots \\ \omega_{n1} & \cdots & \omega_{nn} \end{pmatrix} \tag{3.3}$$

在完成文本向量化的工作后，我们需要继续进行的是对特征项的权重进行计算。

3.2.2 改进 TF-IDF 权重计算法

（1）传统 TF-IDF 权重计算法

在进行特征项权重中的计算时，我们针对本系统为答疑系统，输入的问题一般长度不会过长的特点，因此拟将经过分词工作分出来的词语进行向量化后全部作为文本的特征项进行工作。

在选出的特征项中，我们首先要对其特征项的权重进行计算。在计算特征项权重时，TF-IDF 权重计算法的特征项权重计算公式为：

$$TF\text{-}IDF\ (\omega_i) = TF(\omega_i) \times IDF\ (\omega_i) = n_{i,k} \times \log \frac{N}{N_{i,k} + \alpha} \tag{3.4}$$

在该公式中，虽然已经较为成熟，但是还是存在一定的缺陷：该算法只考虑到了词频和逆文本频率两个因素，但是在实际使用中，有可能出现几个名词的词频相同的情况，在此种情况下，专业名词的权重明显要高于常用名词，但是在该算法中无法体现这一点。基于可能存在不同类型的名词在权重计算法中出现所占权重相同的情况，本文对不同的词汇赋予不同的权重，来进一步精确 TF-IDF 算法计算的结果。

（2）改进 TF-IDF 权重计算法

本文在这里引入权重因子[45]概念：将传统的 TF-IDF 算法计算出的权重 ω_i 赋予权重 α_1，将词汇权重 ω_p 赋予权重 α_2 改进后的权重计算公式为：

$$\omega_{new} = \alpha_1 \times \omega_i + \alpha_2 \times \omega_p \tag{3.5}$$

其中，令 $\alpha_1+\alpha_2=1$，并赋值 $\alpha_1=0.6$，$\alpha_2=0.4$，这样保证传统计算出的权重在新权重计算方法中仍占有较大比重，同时将专业词汇和非专业词汇赋予不同的权重，增加其权重计算的合理性。

其中，ω_p 对于不同词汇的不同赋值情况如表 3.3 所示。

表 3.3 ω_p 不同词汇赋值情况

词类型	权重 ω_p
专业名词	0.6
非专业词汇	0.4

在引入权重因子后，我们可以对权重计算的准确性进行进一步的分析：当“栈”、“队列”和“区别”三个词在经过传统的 TF-IDF 算法计算后，计算的权重 ω_1 均为 0.3 时，此时若直接进行相似度计算，则系统在处理时则会将“栈”、“队列”、“区别”作为同等重要的词汇进行处理，但明显栈、队列的重要程度是要优于区别的。在经过改进 TF-IDF 算法计算后，“栈”和“队列”的权重为：

$$\omega_{new}=\alpha_1\times\omega_i+\alpha_2\times\omega_p=0.6\times0.3+0.4\times0.6=0.42 \quad (3.6)$$

而 “区别”一词的权重则为：

$$\omega_{new}=\alpha_1\times\omega_i+\alpha_2\times\omega_p=0.6\times0.3+0.4\times0.4=0.34 \quad (3.7)$$

通过该例子我们可以看出，本文提出的改进 TF-IDF 权重计算法在一定程度上提高了计算出的权重的准确性，接下来我们将对权重设置的合理性进行进一步分析。

（3）改进 TF-IDF 权重计算法合理性分析

在本部分中，我们将从权重赋值方法、赋值大小的合理性两个方面来对提出的改进 TF-IDF 权重计算法的合理性进行分析。

①权重赋值方法

目前，常用的权重赋值方法包括加法合成法、乘法合成法两种综合权重计算方法。在本文中采用的则是加法合成法对权重进行的计算：将赋权依据不同的主观权重向量 ω_p 和客观权重向量 ω_i 在一定的权重分配下（α_1、α_2）直接相加，该方法在企业规模评级、项目评价指标等方面广泛应用，但在相似度计算方面还未引入；

②赋值大小合理性

将传统算法计算出的客观权重向量 ω_i 赋予 0.6 的权重大小，保证其在改进算法中仍占主导地位；对引入的词性主观权重向量 ω_p 赋予 0.4 的权重大小，保证其在改进算法中有影响但不会在过大差距时影响结果；

对 ω_p 赋予的专业词汇占比 0.6 和非专业词汇占比 0.4，则是根据目前在相似度计算中常用的二因素间的权重分配方式。

3.2.3 问题相似度计算

在计算出各特征项的权重的基础上，需要对其与数据库中的问题进行相似度计算。在本系统中，采用的余弦相似度的计算方法：若计算出的代表两特征项的特征向量之间的夹角越小，则两向量间的余弦值越大，则两向量间的相关性越大，即二者间的相似度越高。在此基础上，文本 T_1，T_2 之间的相似度就计算为：

$$SIM(T_1,T_2)=\frac{\omega_{1k}\times\omega_{2k}}{\|\omega_{1k}\|\|\omega_{2k}\|}=\frac{\sum_{k=1}^{n}\omega_{1k}\times\omega_{2k}}{\sqrt{\left(\sum_{k=1}^{n}\omega_{1k}^{2}\right)\left(\sum_{k=1}^{n}\omega_{2k}^{2}\right)}} \tag{3.8}$$

在式 3.8 中，T_1，T_2 为进行比较的两个文本，ω_{1k}、ω_{2k} 为对应向量 t_1，t_2 的权重。在计算出的结果中，若结果越接近于 1，则说明两向量间的夹角越接近于 0°，则两向量间的相似度越大；若结果越接近于 0，则说明两向量间的夹角越接近于 180°，则两向量间的相似度越小。

3.2.4 答案提取

在完成对问题的相似度计算后，接下来的工作就是从计算出的众多相似的问题中选取部分相应的答案反馈给学生，完成学生的答疑工作。在此工作过程中，主要是通过对比各个问题之间的相似度来选取最佳答案。答案提取的流程大致分为以下两点：

①若相似问题相对较少，则按照系统计算的相似度，按照相似度高低进行排序后，全部反馈给学生，为学生完成答疑工作；

②若相似问题较多，则设定一个阈值（通过实验测试后，该值设置为 10），当相似问题超过该阈值时，则仅选取排名靠前的十个问题为学生进行反馈。

在完成答案提取后，完成了整个自动答疑的过程。自动答疑流程图如图 3.6 所示。

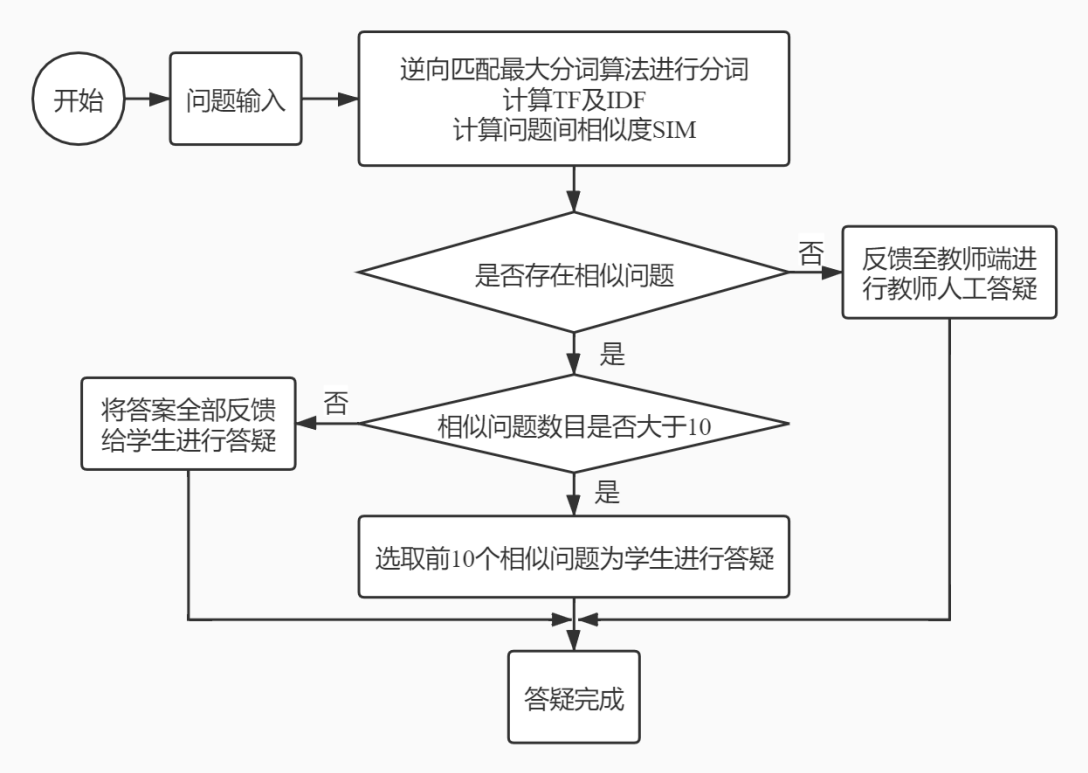

图 3.6 自动答疑流程图

3.3 本章小结

在本章内容中，主要对系统中用到的算法的设计和应用进行了讲述。第一部分讲述的是基于字符串匹配的中文分词方法，首先讲述的是系统中词典的构建，然后是对三种最大匹配算法进行了对比使用，同时对逆向最大匹配分词算法算法的进行了详细的描述；第二部分讲述的是语义相似度的计算，分别介绍了本研究中的改进的 TF-IDF 权重计算方法、问题相似度的计算及问题提取时使用到的方法。

4 基于语义相似度计算的智能答疑系统分析与设计

本文在第 2 章中对基于语义相似度计算的智能答疑系统用到的相关技术及理论进行了介绍，在第 3 章中对研究的智能答疑算法进行了描述，完成了本系统实现的基础。在本章中，将对智能答疑系统的分析与设计做详细的介绍，其中包括需求分析、概要设计、功能模块设计及数据库设计。

4.1 基于语义相似度计算的智能答疑系统需求分析

随着“互联网+”的不断普及和信息技术的飞速发展，各行各业都在借助互联网实现自身的新的发展。其中，“互联网+教育”的模式已经在教育行业慢慢兴起，各种互联网教育、在线教育平台等，但是在网络教育中的答疑不及时、不准确成为了阻碍网络教育的绊脚石。在这种背景下，设计一个智能程度高德答疑系统成为了目前网络教学的一个重要需求。本部分从学生学习需求、教师教学需求及可行性对基于语义相似度计算的智能答疑系统进行需求分析。

4.1.1 学生用户需求分析

在整个学习过程中，答疑始终是促进学生快速掌握知识、快速进步的重要渠道。而也正是因为如此，在传统教育模式中，教师们在教学任务完成后，往往会拿出专门的时间来对学生在本段时间内学习的问题进行答疑，尽量做到问题及时解决，不积累问题，而学生们也正是通过此种形式实现了对知识的掌握和巩固。

而在新兴的互联网教育模式下，由于网络教育所固有的时间不一致性和地理位置的分割性，使得教师及时对学生的问题进行解决存在一定的困难，倘若学生积累问题过多，会使学生在学习的过程中产生对某些问题的断点[46]，直接影响学生学习的积极性和主动性，学习效率将大打折扣。该智能答疑系统将会通过系统自动答疑、系统内学习者相互答疑的方式，在一定程度上解决目前存在的学习者答疑不及时的问题。

结合本系统，学生可以在学习中及时解决存在的一些问题，在系统中准确找到自己问题的答案。本系统提供的学习者一次答疑未能解决的问题，可以后台追问的功能，避免了只是数据库简单搜索的问题；本系统可以提供相似问题的推荐功能，该功能可以更好的帮助学习者从多个角度更好的对知识点进行掌握，学生可以通过答案下的反馈框对问题提出自己的看法，与老师进行交流，及时更正自己在学习上的一些错误，同时该系统提供的学生和教师身份的认证互通机制，一

方面可以帮助学生更好的解决问题，另一方面可以充分的激发学生学习的积极性。综上所属，从学习者学习的角度来看，存在对该系统的需求。

4.1.2 教师用户需求分析

在教师的教学过程中，学生的反馈对于教师的课程安排、教学进度有着至关重要的作用[47]。在传统教育中，教师通过与学生之间的直接面对面交流，可以及时准确的了解到学生对知识的掌握程度、疑难点所在，根据学生的反馈及时调整教学计划，保证教学的顺利进行。但是在网络教育中，教师无法直接面对每一位同学，不能看到每一个学生的上课听课状况、学习状况和做题情况，对自己课程的安排只能依靠自己经验，存在一定的误区。

在本系统中，教师主要有两部分成员构成：在校的任课老师及通过身份认证、具有成为为学生答疑者能力的用户。对于在职教师来说，该系统可以使其在课后及时为学生答疑，解决学生在学习过程中出现的问题；同时，教师可以通过登录本系统可以了解到学生对那些问题搜索频率高、哪些知识点出现问题多来及时调整自己的教学计划，完善教学过程。对于非在职教师来说，大部分为对该课程兴趣较高、或者是对本课程掌握较好的在校学生，他们都是课程的学习者，通过设立教师身份的认证机制，可以充分激发其学习的主动性，主动学习更多的知识，更好地促进学生的学习。因此从教师用户的角度来看，存在对本系统的需求。

4.1.3 可行性分析

在技术方面，基于语义相似度计算的智能答疑系统主要采用 Python 作为开发的语言来对系统进行部署。在数据库获取中使用了网络爬虫技术，可以直接获取构建数据库所需的数据；在对输入的问题进行分析时采用自动分词技术、语义相似度计算中的向量空间模型等技术，能够准确的对问题进行分析理解，实现与数据库中的精准匹配与相似问题的推荐；

在经济方面，系统所需的问题数据库可以在书籍、网络上免费获取，Python 中使用 Pycharm 对软件进行书写，在经济方面同时也具有可行性。

4.2 基于语义相似度计算的智能答疑系统概要设计

本文要研究的基于语义相似度计算的智能答疑系统是以传统的答疑系统形式为基础，在此基础上添加学习者互答、教师解答、系统相似问题推荐等功能的智能答疑系统。该系统主要可分为用户与管理员两大模块，系统的主要功能示意图如图 4.1 所示。

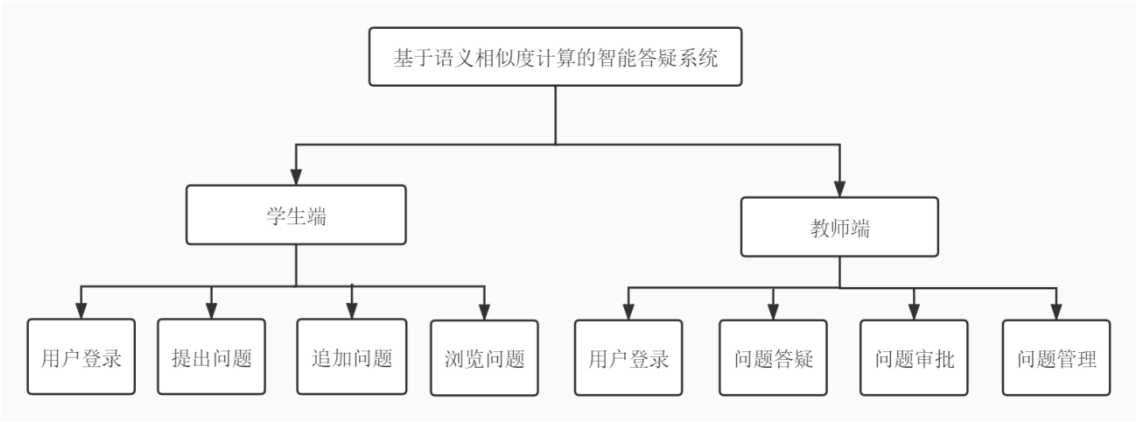

图 4.1 系统主要功能示意图

用户模块下又可具体分为自动答疑模块和浏览其他问题模块，其中自动答疑模块是本系统的核心所在，其主要是采用的是基于字符串匹配的方法和向量空间模型算法。在用户模块中，当用户输入问题时，系统首先对其进行的是字符串匹配方法对问题进行处理，在处理完成后使用的是向量空间模型对其相似度进行计算，在计算过程中设定一个最小的相似度值，若输入问题与数据库中的问题相似度大于该值，则将该问题反馈给学生进行答疑，若小于该相似度，则判定系统中无该问题，将问题转交至教师端，由教师进行人工答疑。

管理员模块主要分为系统管理和教师维护两部分，其中系统管理为后台管理员对系统进行日常维护、用户管理和数据库管理三部分，系统日常维护主要是解决系统功能上的问题或者出现的漏洞，用户管理则是对该系统中的所有注册用户进行管理，定时清理一些长期未登录或在系统中无正常作用的用户，数据库管理是对系统中的数据库进行管理和更新；教师维护模块主要包括问题答疑和问题管理两个功能模块，其中问题答疑是对系统中学习者存在的问题但数据库中并未存在答案的问题进行解决；问题管理是对系统中存在的未解决的问题进行分类管理，可以删除一些与本课程无关的问题。

4.3 基于语义相似度计算的智能答疑系统功能模块设计

本节主要对基于语义相似度计算的智能答疑系统中个各个功能模块的设计进行概述，其中包括：用户注册登录模块设计、学生用户模块设计、教师用户模块设计三个部分。

4.3.1 用户注册登录模块设计

在使用本系统系统之前，使用者需要在本系统中进行注册用户后方可以使用。用户注册过程是在系统中点击注册按钮，进入注册页面。在注册过程中，用户需

要输入用户名、密码等相关信息，当系统验证无重名用户且密码强度满足要求时，则显示注册成功且将该信息添加到用户数据库中进行保存，用户可用该账号进行登录。用户注册登录流程图如图 4.2、图 4.3 所示。

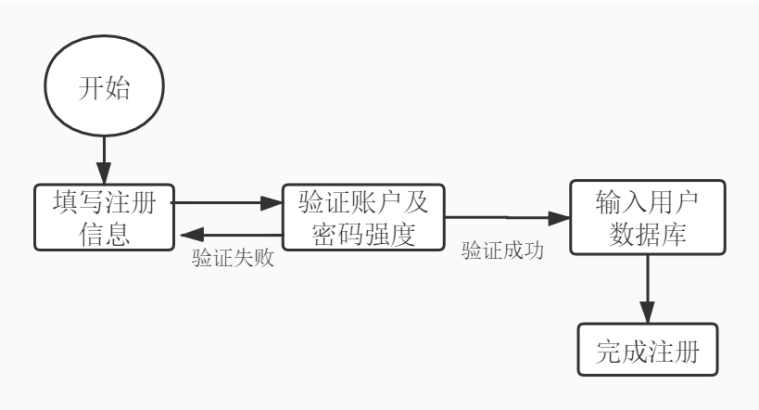

图 4.2 用户注册流程图

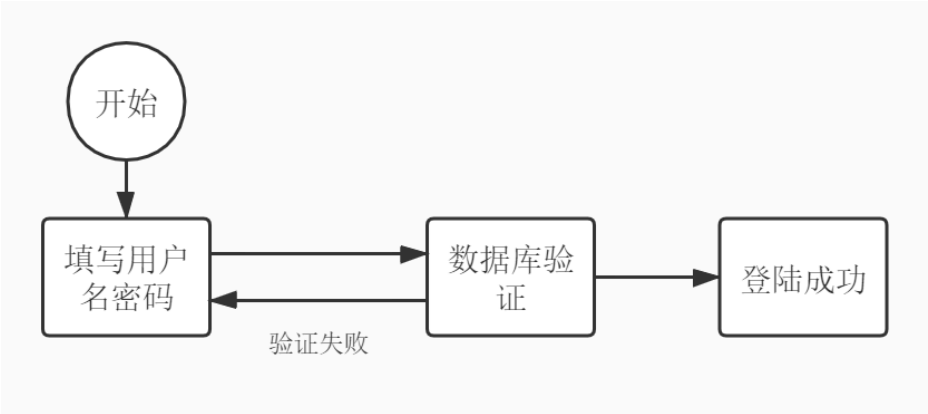

图 4.3 用户登录流程图

4.3.2 学生用户模块设计

在学生用户模块中，主要包括为自动答疑、浏览其他问题两个子功能模块。

自动答疑为整个系统中的核心部分，学生用户的大部分问题均由系统自动答疑来完成。该过程具有准确高效的特点，在自动答疑中，主要的流程包括以下几个方面：

①学生用户成功登陆本系统，在系统上方的问题框中输入问题；

②输入完成后，点击搜索，将问题提交至后台数据库中；

③学生用户将问题提交后，系统将从后台数据库中对问题进行分词、相似度计算等处理，来匹配相似度最高的问题；

④匹配成功后，将问题答案反馈至学生界面；

⑤若未匹配到相似问题，则将该问题提交至教师端，并反馈给学生“暂无答案，已提交至教师处理”；

⑥学生若对问题仍有疑问或者有不同的想法，可以在追加框内输入自己的想法进行再次提问或者与老师进行交流，追加框内处理过程与问题处理过程一致；

⑦当老师为学生回答数据库中不存在的问题时，系统会自动反馈至学生用户，当学生用户再次登录本系统时，系统会给出提示“教师已答疑上次问题”。

整个自动答疑的流程如图 4.4 所示。

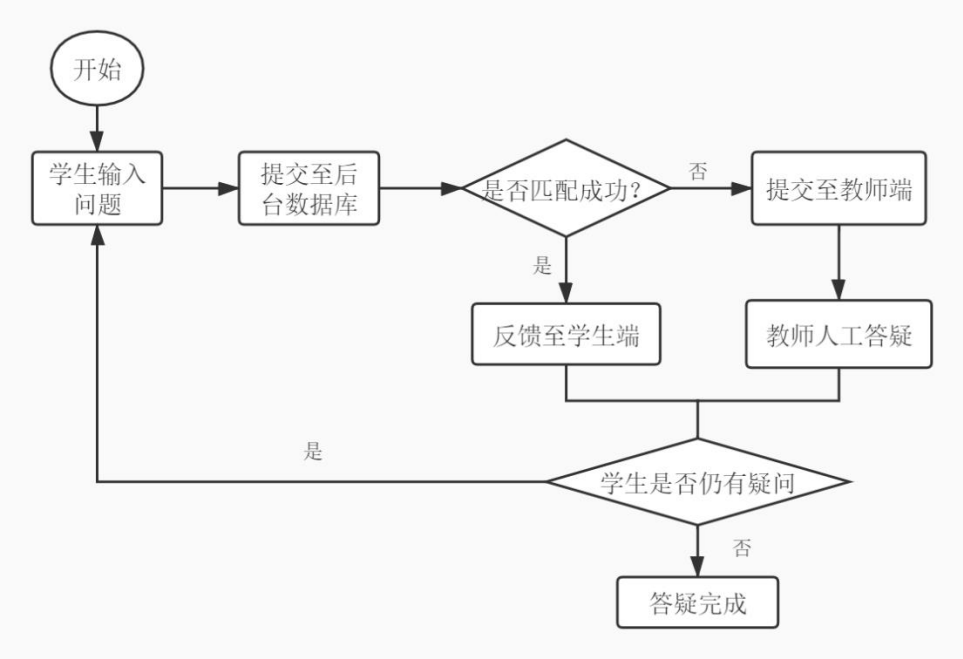

图 4.4 自动答疑流程图

在浏览问题子模块中，主要的功能是对用户搜索较为集中的问题以及对人工答疑过程中教师给予的问题答疑进行浏览，该模块的设计主要是为了学生用户通过多角度对课程进行学习，能够对课程有个更深刻具体的认识。

4.3.3 教师用户模块设计

在教师用户模块中，主要功能包括：问题答疑和问题管理。

在问题答疑阶段，主要是对学生提出的但是数据库中还并未包含的问题进行人工解答。该功能的主要流程为：

①教师登录系统的教师端，查看学生提交的存在的问题；

②教师对问题进行解答；

③在教师解答完成后，将问题及答案反馈给学生，来帮助学生完成答疑；

④系统在教师答疑完成后，自动将该问题及答案收录到数据库中，完成数据

库的更新。

整个问题答疑的流程图如图 4.5 所示。

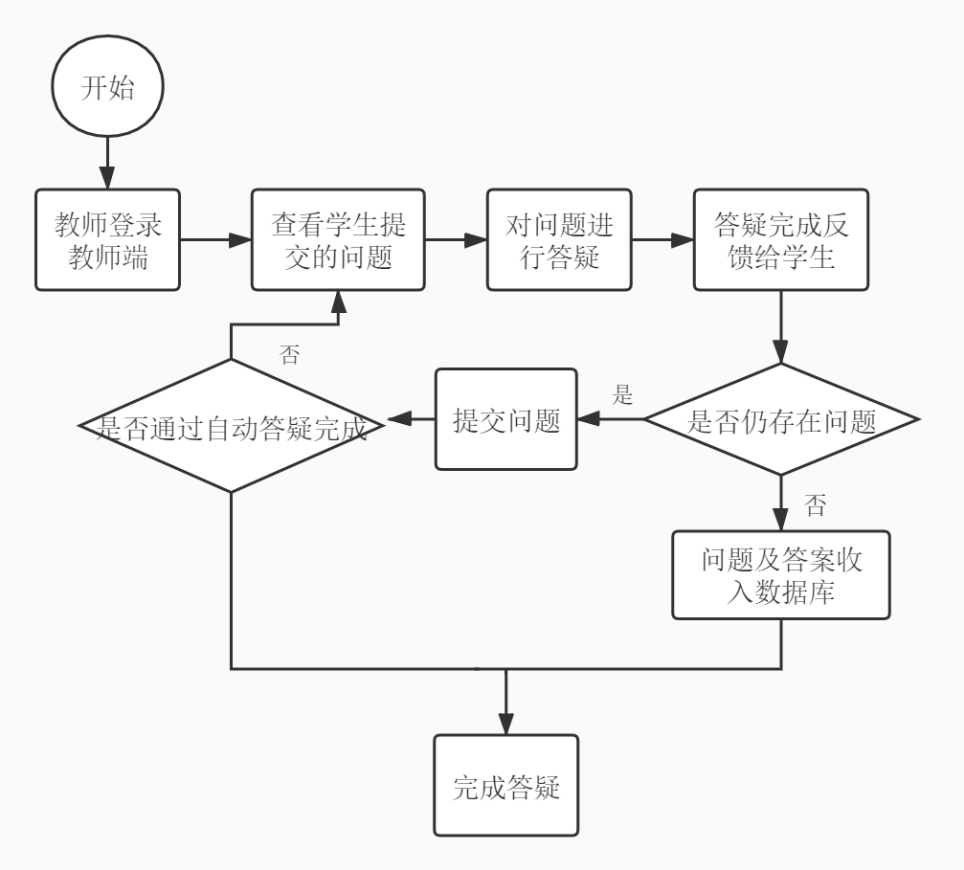

图 4.5 教师答疑流程图

在问题管理阶段，教师主要是对输入的问题进行查看管理，可以查询系统中查询频率最高的问题、哪些知识点学生的追加问题较多，也可以删除一些与本课程无关的问题或者添加一些教学上出现的新问题到数据库中来供学生学习和为学生进行答疑。

4.4 数据库设计

在每一个系统中，数据库都是其开发的核心和基础[48]。而数据库的设计，就是为软件建立一个能够与其功能相匹配的数据库，在整个系统软件部署过程中有着极其重要的作用。如果数据库的设计出现问题，会导致系统中出现一系列问题，不利于系统的稳定性。

数据库设计[49]主要可以分为数据库的逻辑设计和物理设计。在数据库的逻辑设计[50]中，就是把概念结构设计阶段设计好的基本实体-关系图转换为与选用的数据库管理系统产品所支持的数据模型相符合的逻辑结构；而数据库的物理设计就

是为数据库逻辑设计的逻辑结构寻找一个具体的工作环境，包括：数据类型、实体属性、DBMS 页面大小等。在对数据库进行设计时，要遵循一对一设计原则、独特命名原则、双向使用原则三个原则进行设计。

在本系统中，采用 MySql 数据库存储数据的相关信息。在本系统中，共有三类用户会对本系统的数据库进行使用，分别为学生用户、教师用户及管理员用户。其中，学生用户是在提交问题及问题反馈时，与数据库中的问题信息表和答案信息表进行交互；教师用户是在输入新问题、为学生答疑后新问题输入数据库中与数据库的问题信息表和答案信息表进行交互，对数据库中不存在的问题进行答疑时会与教师答疑信息表进行交互；管理员用户是对整个系统中的用户进行管理，会通过用户信息表对学生用户和教师用户进行操作。

根据本系统对数据库的需求，在系统中设计的数据表及其数据项和数据结构如下所示：

①用户信息表（user）：主要用来存储用户的相关信息，包括用户 ID、用户名、用户密码、注册时间、登陆时间、用户权限等。

表 4.1 用户信息表

字段名称	存储信息	数据类型（长度）	允许空	是否主键
UID	用户 ID	INT(10)	否	是
Uname	用户名	Varchar(100)	否	否
Upassword	用户密码	Varchar(20)	否	否
Uregeade	注册时间	DateTime(8)	否	否
Ulogin	登陆时间	DateTime(8)	否	否
Urights	用户权限	INT(10)	是	否

②问题信息表（question）：问题信息表是确保答疑准确性数据表，其中主要存储了数据库中的问题、问题的编号、问题关键词、问题答疑次数、提问该问题的学生用户等信息。

表 4.2 问题信息表

字段名称	存储信息	数据类型（长度）	允许空	是否主键
QID	问题编号	INT(10)	否	是
Question	问题全文	Varchar(1000)	否	否
Qkeyword	关键词	Varchar(100)	否	否
Qcount	答疑次数	INT(10)	是	否
StudentID	提问学生	Varchar(50)	是	否

③答案信息表（answer）：答案信息表主要用来存储与问题一一对应的答案的

相关信息。其中主要存储了：问题编号、问题答案、访问次数、修改次数、修改时间、修改人员等信息。

表 4.3 答案信息表

字段名称	存储信息	数据类型（长度）	允许空	是否主键
QID	问题编号	INT(10)	否	是
Answer	问题答案	Varchar(1000)	否	否
Acounts	访问次数	INT(10)	是	否
Arevise	修改次数	INT(10)	是	否
Areperson	修改人员	Varchar(50)	是	否
Aretime	修改时间	DateTime(8)	是	否

④教师答疑信息表（teaanswer）：教师答疑信息表主要是用来存储学生提问时，数据库中未找到相似问题需要教师进行人工答疑时，问题存放的信息表。其中主要存储了：问题编号、问题全文、待答疑者、提问时间、是否答疑、答疑内容提供、答疑教师、答疑时间等。

表 4.4 教师答疑信息表

字段名称	存储信息	数据类型	允许空	是否主键
QID	问题编号	INT(10)	否	是
Question	问题全文	Varchar(100)	否	否
StudentID	待答疑者	Varchar(50)	否	否
StudentTime	提问时间	DateTime(8)	否	否
State	答疑状态	Varchar(2)	否	否
Answer	答疑内容	Varchar(1000)	否	否
TeacherID	答疑教师	Varchar(50)	否	否
TeacherTime	答疑时间	DateTime(8)	否	否

4.5 本章小结

本章为基于语义相似度计算的智能答疑系统分析与设计，在本章的第一部分对本系统进行了需求分析，其中包括学生用户需求分析、教师用户需求分析及系统的可行性分析，从各个用户的角度对系统进行了分析；第二部分对系统进行了概要设计，对整个系统的流程进行了简单的描述；第三部分是系统的模块设计，分为用户注册登录模块设计、学生用户模块设计和教师用户模块设计，文中对各个模块的具体流程进行了讲解；第四部分是对系统的数据库设计进行了描述，对各个数据表及数据项和数据结构进行了描述。

5 基于语义相似度计算的智能答疑系统的实现与测试

本章节将针对基于语义相似度计算的智能答疑系统的系统功能模块的测试与系统性能的测试两个部分，来对本系统进行进一步阐述。

5.1 系统开发环境

该系统目前设计的主要是 PC 端使用，使用语言为 Python 语言，计算机系统为 Windows 系统。本系统测试的具体环境如表 5.1、表 5.2 所示：

表 5.1 系统测试硬件环境

系统硬件	硬件环境
机器型号	联想 Y430P
CPU	Inter Core i5
内存	8G
硬盘	500G

表 5.2 系统测试软件环境

软件名称	软件版本
Windows	Windows 10
Python	Python 3.7
Mysql	5.1

5.2 基于语义相似度计算的智能答疑系统基本功能实现

在本系统中包含许多功能，本部分将针对注册新用户、自动答疑、教师答疑、后台用户管理四个方面对系统已实现的基本功能进行描述。

5.2.1 注册新用户

系统中的用户共分为三类：答疑者（学生）、教师和管理员。在用户首次使用本系统进行答疑时时，需要注册一个学生端个人账号密码来登录本系统，学生端的用户注册是开放的，任何人在保证密码强度前提下均可注册，注册后即可登录系统进行提问；教师端的注册权限需要使用者备注提交个人信息，在后台管理员对用户提交信息进行审核后，若该用户具有成为教师端进行答疑的能力，则将其批准成为教师端，可以对用户的问题进行直接答疑，无需后台教师端审核；管理员用户不允许注册。注册新用户登录界面如图 5.1 所示。

图 5.1 用户注册登录界面

5.2.2 自动答疑

用户登陆本系统后，会直接看到系统中的问题提问框与问题答疑框。学生在将问题输入后，系统会自动对问题进行分析处理，处理完成后与数据库中的数据进行分析比对，在完成这一过程后为学生的问题进行答疑，若在数据库中检测到相似度 95%以上的问题，则对问题进行直接答疑，同时出现多个问题满足时，则按顺序逐一显示，学生可以点击下一条进行查看；若数据库中不存在满足该相似度的问题，系统则将该问题自动提交到教师端进行人工答疑，同时返回给学生“暂时无人作答，已提交到教师端”的消息。学生在接受到答案的同时，还可以继续对问题进行追问或者提出自己的想法提交到教师端，在教师端对学生提交问题进行处理后，可将正确的想法思路添加到数据库中，完善数据库资源。问题答疑界面如图 5.2 所示。

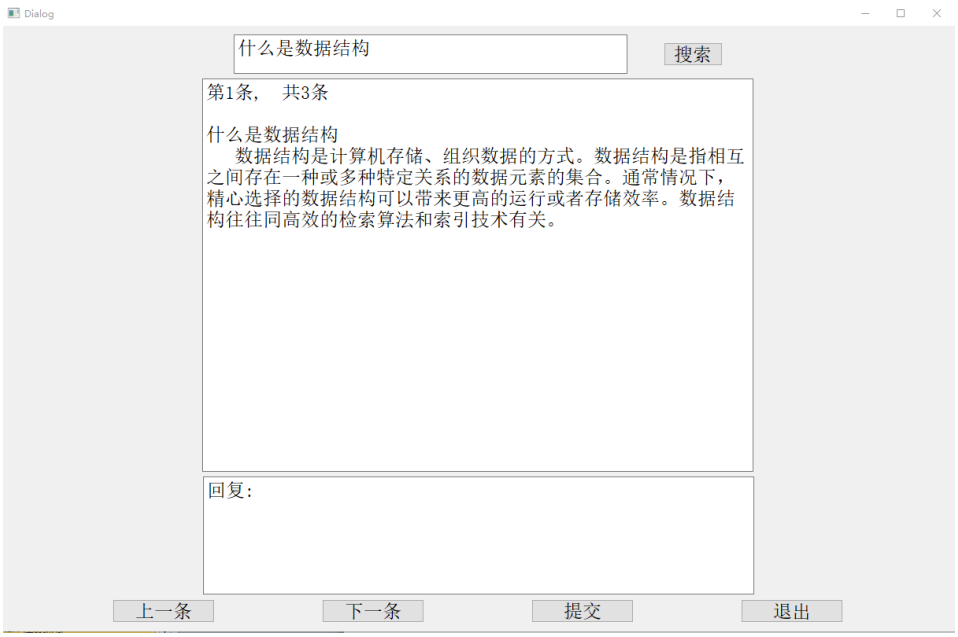

图 5.2 学生答疑界面

5.2.3 教师答疑

当学生提交的问题数据库中匹配不到答案时，问题提交到教师端由教师进行答疑，在教师答疑端，主要有两个功能：一是可以对学生提出的问题进行直接答疑，答疑后系统将会把老师的答案反馈给学生，对学生的提问进行答疑；二是可以对学生提出的想法进行批阅，若同意学生对问题的解答，点击同意则将学生的答案输入数据库中，完善该问题的答案。教师答疑界面如图 5.3 所示。

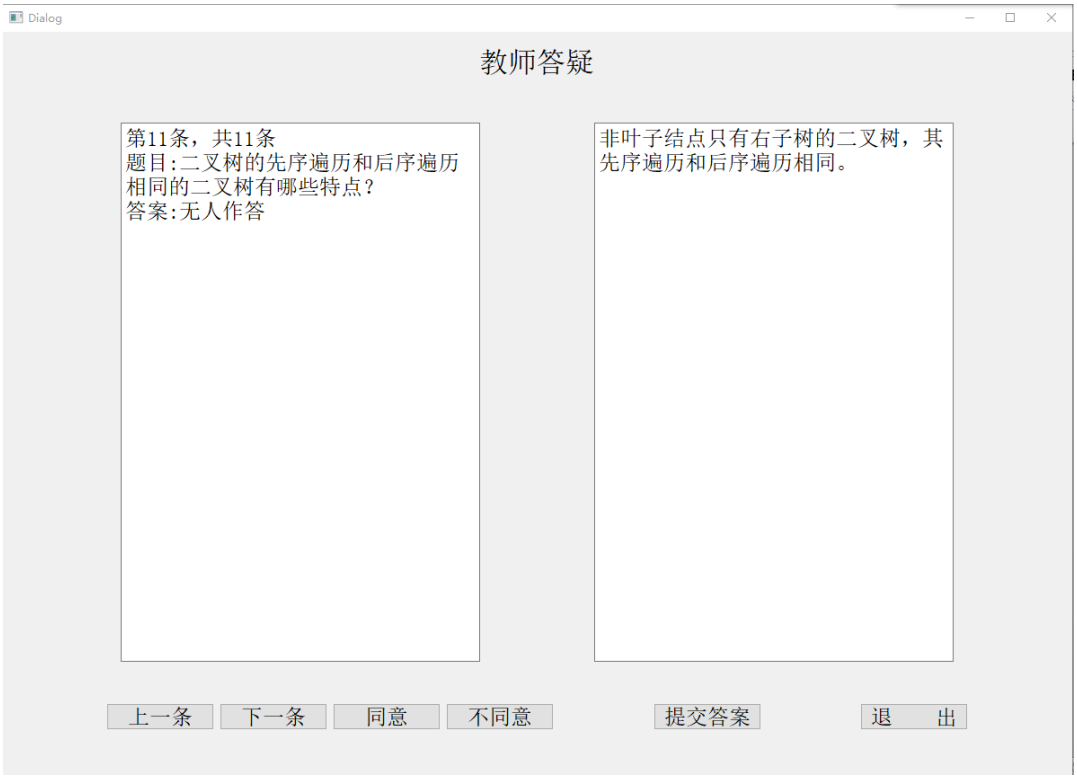

图 5.3 教师答疑界面

5.2.4 后台用户管理

在后台管理员界面中，主要的功能是对申请成为教师端的用户进行审批，管理员可以根据用户提供的个人信息对用户进行评审，若该用户满足成为教师端的要求，则通过其申请，让其可以为其他学生的问题进行答疑；若该用户不满足要求，则拒绝其请求，让其只能是学生端用户。后台用户管理界面如图 5.4 所示。

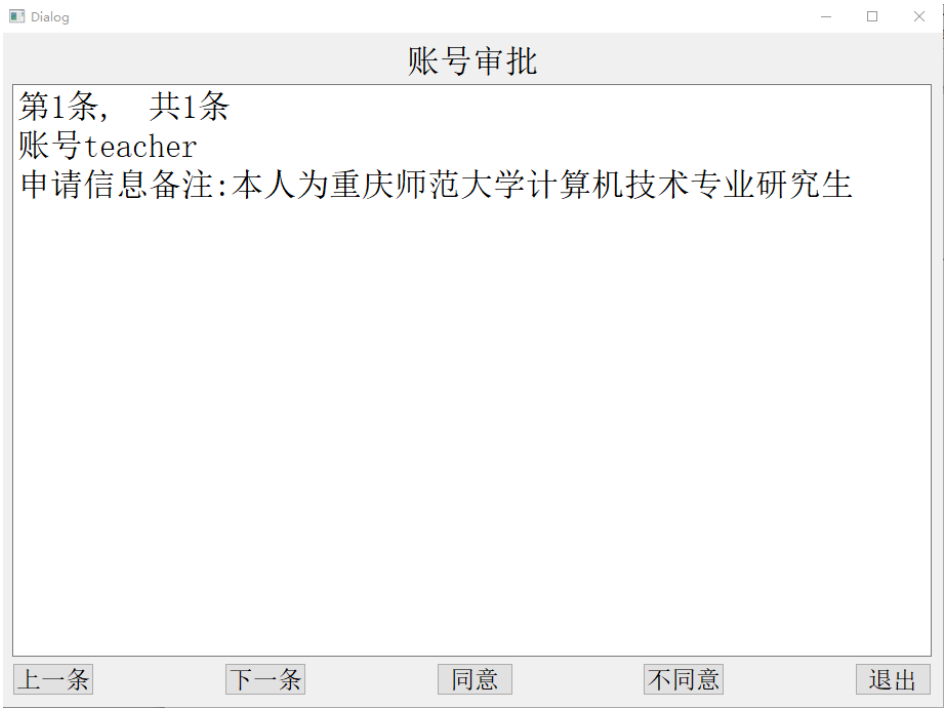

图 5.4 后台管理员审批

5.3 基于语义相似度计算的智能答疑系统测试

在完成系统基本功能的实现后，进行对基于语义相似度计算的智能答疑系统的测试，来检测系统能否正常使用。在测试的过程中，主要采用的是单元测试的方法。

5.3.1 系统测试目的

进行答疑系统的测试，主要是为了保证本系统的准确性和高效性，测试过程和结果并不能说明本系统已经完善，只能说明本系统已经在一定程度上完成了预期该程序设计的功能。在后续的实验中，还将会对目前系统存在的不足进行进一步的更新完善。

通过测试，目前本系统已实现的功能大致有以下几点：

①能够对输入的问题使用基于字符串匹配的方法进行准确分词，保证问题在输入数据库进行答疑时的精准匹配；

②能够在问题进入后台问题库时，使用 VSM 模型对输入问题与问题库中相关问题进行对比，随后输出与输入问题相似度最高的问题的答案；

③能够在学生提出系统未收录问题时，对学生进行系统中已有相似问题的推荐。

5.3.2 基于语义相似度计算的智能答疑的单元测试

本节将会针对本系统中的每一功能模块分别进行单元测试，测试包括用户注册登录模块、学生模块、教师模块、相似问题推荐模块等等。各模块测试结果如表 5.1-表 5.4 所示。

表 5.1 用户注册登录模块测试及结果

序号	测试功能	测试方法	预期结果	实际结果
1	学生用户注册	正确填写各项信息	跳转至后台用户注册页面并注册成功	跳转至后台用户注册页面并注册成功
2	学生用户登录	帐号密码输入正确	登录成功	登录成功
3	学生用户登录	帐号密码输入错误	提示密码错误	提示密码错误
4	教师用户注册	填写个人信息	提示填写备注信息并移交管理员	提示填写备注信息并移交管理员
5	教师用户登录	帐号密码输入正确	登录成功	登录成功
6	教师用户登录	帐号密码输入错误	提示密码错误	提示密码错误
7	管理员用户注册	点击管理员注册	显示不允许注册	显示不允许注册
8	管理员用户登录	帐号密码输入正确	登陆成功	登陆成功

表 5.2 学生提问模块测试及结果

序号	测试功能	测试方法	预期结果	实际结果
1	学生提问	输入问题	进行答疑	进行答疑
2	学生追加问题	反馈框内追加问题	提交教师答疑	提交教师答疑
3	学生对问题答案补充	反馈框内补充答案	提交教师审批	提交教师审批
4	学生提问数据库中尚未收录的问题	输入问题	系统提交给就教师端，反馈尚未收录	系统提交给就教师端，反馈尚未收录

表 5.3 教师答疑模块测试及结果

序号	测试功能	测试方法	预期结果	实际结果
1	教师进行答疑数据库中不存在的问题	提问未收录问题后，对未收录问题进行答疑	答疑后答案收录到数据库中	答疑后答案收录到数据库中
2	教师审批学生的答案补充	学生端提交后，教师端进行操作	审批通过录入数据库，不通过反馈给学生原因	审批通过录入数据库，不通过反馈给学生原因
3	教师补充答疑学生对答案的疑问	学生端完成答疑后追问，教师端进行操作	教师答疑后反馈至学生端	教师答疑后反馈至学生端

表 5.4 相似问题推荐模块测试及结果

序号	测试功能	测试方法	预期结果	实际结果
1	已收录问题相似问题推荐	学生提出问题，查看推荐相似问题	相似问题基本能满足学生补充练习需求	相似问题基本能满足学生补充练习需求
2	未收录问题相似问题推荐	学生提出未收录问题时，查看相似问题推荐	能在一定程度上解决学生的问题	能在一定程度上解决学生的问题

5.3.3 系统测试结果

在本系统中，最为核心的部分为自动答疑部分，因此在测试过程中，主要针对答疑的准确性和答疑效率方面进行了测试。在测试过程中，将数据库中未含有、提交教师人工答疑的问题统一作为非正确答疑，经部分学生测试后，得到的测试结果如表 5.5 所示：

表 5.5 系统准确率测试结果

学生 ID	问题数量	正确答疑数量	准确率	主要提问的问题类型
001	50	45	90%	定义类问题
002	35	32	91%	定义类问题
003	21	20	95%	原因类问题
004	44	40	91%	原因类问题
005	18	16	89%	计算类问题
006	31	26	84%	计算类问题
007	26	23	88%	方法类问题
008	39	33	85%	方法类问题
平均准确率			89%	

表 5.6 系统答疑效率测试结果

问题编号	问题	答疑时间（ms）	答疑结果
1	什么是线性表	6	正确
2	二叉树有什么特点	8	正确
3	栈和队列的共同点	7	正确
4	常见排序算法有哪些	4	正确
5	图的遍历有哪些方法	62	正确
6	无向图的邻接矩阵是什么	31	正确

（续表 5.6）

7	高度为 5 的完全二叉树最多多少结点	90	未录入数据库
8	什么是深度优先遍历	11	正确
平均效率		27	

从测试结果中可以看出，本系统的平准准确率为 89%，由于初步的数据库并不完善且提交教师答疑的问题在本测试中均作为正确答疑处理，因此当本系统投入使用后准确率还能有进一步提升。当用户提问定义类、原因类问题时，系统能够以 90%以上的正确率进行答疑，能够基本满足学生对定义类问题的需求，但是当用户提问计算类、方法类问题时，系统的正确率略有下降，分析是由于计算类、方法类问题数字容易造成检索时的重复，面对这种问题时，学生可采取多输入文字的方式进行答疑，尽量避免大量数字的出现。

在答疑效率方面，本系统的平均效率值为 27ms，由于存在数据库中不存在的问题，因此检索时间会随系统的使用和数据库的完善逐渐减少，可以保证后续效率值逐步提升。从性能测试的结果来看，本系统的性能已经基本上可以满足学生进行日常答疑的需求。

5.4 本章小结

本章主要对基于语义相似度计算的智能答疑系统的实现与测试进行了概述。首先对系统的开发环境中的硬件环境和软件环境进行了说明，然后对系统的基本功能进行了概述，最后对系统整体进行了测试，分别对对系统的各个模块进行了单元测试、对系统整体进行了整体测试。总体来说，系统的各模块功能与预期功能基本一致，系统整体性能表现良好，已可基本满足学生的日常答疑需求。

6 总结和展望

6.1 论文工作总结

本文完成的主要工作就是在语义相似度计算的理论基础上，设计搭建了一个准确率较高、答疑更为准确智能的智能答疑系统。本系统相较于现有的答疑系统，综合了传统搜索引擎答疑和在线教师答疑的两种方式的优点，能够帮助学生在最大程度上完成答疑的相关工作，其性能也具有较好的数据反应。

本文完成的主要工作有以下几点：

（1）对智能答疑系统和语义相似度计算的研究现状进行了分析。在目前研究下，答疑系统和语义相似度计算都已经具有较好的研究基础，可以直接进行系统的开发和研究；

（2）对本文使用到的相关技术及理论进行了概述。分析对比了目前常用的三种语义相似度计算方法、三种中文分词技术和四种权重计算方法；

（3）对本文使用到的算法的设计和应用进行了讲解。主要对比研究了三种常用中文分词算法，同时构建了系统词典，讲述了逆向最大匹配算法的算法流程，语义相似度的计算中使用到的算法及其流程，提出了改进 TF-IDF 权重计算法；

（4）对本系统的分析与设计。首先是分两种身份对本系统的不同使用者的需求进行了分析，同时对系统的概念设计和三个功能模块的设计进行了解释，最后，设计了系统使用到的数据库；

（5）对本系统的实现及测试。对系统的开发环境进行了说明，同时展示了系统四个常用界面后，对后期进行的系统测试的结果进行了表述。

6.2 下一步工作展望

本研究在设计与搭建智能答疑系统方面有了一定的成果，但是个人认为，本系统在以下几方面还存在可以继续改进完善的地方：

（1）语音识别系统的引进。目前来说，语音识别系统在各个行业都已经取得不小的成绩，但是在目前的答疑系统中，还并没有较为成功的案例出现。在本系统设计初期本想引入语音识别功能，但是由语音识别算法需要大量的语音数据库对算法进行训练，在本研究中暂未引入；

（2）加入图片识别功能。在本系统测试期间，发现部分较为复杂的数学公式，在输入计算机的过程中存在一定的困难和误差，通过了解后发现，图片识别领域的文字识别目前已经有较好的成果和产品，由于缺乏专项数学公式数据集的训练，

在本研究中也暂未引入，在日后的工作中获得完备的数学公式数据集后，将在研究中增添该功能，由系统完成由图片到文字的转换，实现更为高效的答疑；

（3）系统界面的优化。在本系统中的系统界面相对较为简洁，准备在日后的工作中对系统的界面进行进一步的优化，提升学生、教师使用时的舒适感，提高系统的人性化设计。

参考文献

[1]杨苏琴,张晋峰.基于现代教育技术与高校体育课堂教学整合的研究思考[J].经济师. 2019(12): 209-210.

[2]Kuo Y, Walker A E, Belland B R, et al.A Predictive Study of Student Satisfaction in Online Education Programs[J].International Review of Research in Open and Distributed Learning. 2013, 14(1): 16-39.

[3]马新意.自动答疑系统中文分词模块的设计与实现[J].信息技术与信息化 2019,19(01):22-25.

[4]李印鹏.教学网站智能答疑系统设计与实现[D]. 河北科技大学, 2019.

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

[6]王东升,王为民,王石,等.面向限定领域问答系统的自然语言理解方法综述[J]. 计算机科学. 2017, 44(8): 1-8, 41.

[7]段昊昱.慕课平台上的智能答疑系统的设计与实现[D].天津师范大学, 2018.

[8]Richard S,Wallanc.ALICE Primary resource[EB/OL].http://aliee.sunlitsurf.c om/aliee/about.html, 2003-12-26/2010-12-10. Ask Jeeves.

[9]Elhalwany I, Mohammed A, Wassif K, et al.Using Textual Case-based Reasoning in Intelligent Fatawa QA System[J].The International Arab Journal of Information Technology.2015,13(5): 201-210.

[10]蒲黎明.电信诈骗语义分类系统的设计与实现[D].北京邮电大学, 2019.

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

[12]Boris Kate, Gregory Marton, Gary Borchardt, et al. The START Natural Language Question Answering system[EB/OL]. http://start. csail. mit. edu, 2006-2-12/2010-12-10.

[13]Hao T,Qiu X,Jiang S.Leveraging Semantic Labeling for Question Matching to Facilitate Question-answer Archive Reuse//Internati onal Conference on Intelligent Computing[Z]. 2015：65-75.

[14]陈丽,李爽.国内外网上智能答疑系统比较研究[J].中国电化教育. 2003(192).

[15]周睿斌,尤晋元,申瑞民.基于WWW的远程教学中Answer Web的建立[J].计算机工程与应用. 1998(12): 21-23.

[16]丰乃波.以学习者为中心的远程智能答疑系统的设计与实现[D].江西科技师范大学, 2016.

[17] Pukkaew C. Assessment of the Effectiveness of Internet-Based Distance Learning through the VClass e-Education Platform[J]. International Review of Research in Open and Distributed Learning. 2013, 14(4): 255-276.

[18]张波.基于维基百科链接特征的词语语义相似度计算[J].软件工程. 2019, 22(10): 36-43.

[19] Gali N, Mariescu-Istodor R, Hostettler D. Framework for syntactic string similarity measures[J]. 2019, 129: 169-185.

[20]王春柳,杨永辉,邓霏,等.文本相似度计算方法研究综述[J].情报科学. 2019, 37(03): 158-168.

[21]杜坤,刘怀亮,郭路杰.结合复杂网络的特征权重改进算法研究[J].现代图书情报技术. 2015(11): 26-32.

[22]欧阳林艳.VSM在旅游自动问答系统中的应用研究[J].山西能源学院学报.2019, 32(2): 97-99.

[23]冯高磊.基于VSM结合词语语义的文本相似度算法研究[D].北京建筑大学, 2018.

[24]曾文.基于VSM的科技期刊文献与专利文献的相似度计算方法研究[J].情报工程. 2016, 2(3): 37-42.

[25] M. R. Comparing Boolean and probabilistic Information Retneval Systems Across Queries and Disciplines [J]. the American Society for In formation Science. 1997, 48(2): 143-156.

[26] Kuncheva L I. Fitness functions in editing K-NN reference set by genetic algorithms[J]. Pattern Recognition. 1997, 30(6): 1041-1049.

[27]吴代文,杨方琦. Lucene在数据库全文检索中的性能研究[J]. 微计算机应用. 2011, 32(6): 53.

[28]杨涛.中文信息处理中的自动分词方法研究[J].现代交际,2019(07):93-95.

[29]韩冬煦,常宝宝.中文分词模型的领域适应性方法[J].计算机学报. 2015,38(02): 272-281.

[30]胡锡衡.正向最大匹配法在中文分词技术中的应用[J].鞍山师范学院学报. 2008(02): 42-45.

[31]丁振国,张卓,黎靖.基于Hash结构的逆向最大匹配分词算法的改进[J].计算机工程与设计. 2008(12): 3208-3211.

[32]麦范金,王挺.基于双向最大匹配和HMM的分词消歧模型[J].现代图书情报技术. 2008(08): 37-41.

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

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

[35]孔振.基于VSM的文本分类系统的设计和实现[D].哈尔滨工业大学,2014.

[36]Wang D, Zhang H, Liu R, et al.t-Test feature selection approach based on term frequency for text categorization[J]. Pattern Recognition Letters. 2014, 45: 1-10.

[37]Azam N, Yao J. Comparison of term frequency and document frequency based feature selection metrics in text categorization[J]. Expert Systems with Applications. 2012, 39(5): 4760-4768.

[38]Peng T, Liu L, Zuo W.PU text classification enhanced by term frequency-inverse document frequency-improved weighting[J]. Concurrency and Computation: Practice and Experience.

2014, 26(3): 728-741.

[39]杨楷.基于信息熵的权重计算方法在隐含狄利克雷分布中的探索及研究[D].华南理工大学, 2017.

[40]甘秋云.基于TF-IDF向量空间模型文本相似度算法的分析[J].池州学院学报.2018,32(03): 41-43.

[41]李素建.基于语义计算的语句相关度研究[J].计算机工程与应用.2002(07): 75-76.

[42]蒋卫丽,陈振华,邵党国,等.基于领域词典的动态规划分词算法[J].南京理工大学学报,2019，43(1):63-71.

[43]郑国兴.面向航天领域的中文分词算法研究与实现[D].西安电子科技大学, 2019.

[44]郑木刚,刘木林,沈昱明.一种基于词典的中文分词改进算法[J].软件导刊,2016,15(3):42-44.

[45]张振峰.基于向量空间模型的文本分类算法研究[D].杭州电子科技大学, 2012.

[46]周鹏超,王兴辉.远程教育学习者学习成绩影响因素及解决途径[J].辽宁广播电视大学学报,2020(1):41-43.

[47]马贵平.信息技术与教师教学方式的有效整合[J].信息技术与教学,2019(35):138.

[48]李俊梅.计算机软件开发中的数据库测试技术探讨[J].中国新通信,2019,9(21):159-160.

[49]赵锡娟.基于Asp.net的数据库技术基础教学平台的设计与实现[J].电脑知识与技术.2011, 7(30): 7338-7339.

[50]周艳平,李金鹏,蔡素.基于同义词词林的句子语义相似度方法及其在问答系统中的应用[J].计算机应用与软件. 2019, 36(08): 65-68.

附录 A：作者攻读硕士学位期间发表论文及科研情况

[1] 邢政. 计算机便携式输入板，实用新型专利一项

致谢

光阴似箭，日月如梭。转眼间，两年的研究生学习生涯已经走到了最后。回想开学那天还恍如昨日，现在却已走到了毕业的十字路口。

我首先想要感谢的，是我的研究生导师，魏延教授。魏老师是我在学业路上的领航人，是我在生活中学习的榜样。无论我在学习生活中遇到点滴困难，或是论文撰写过程中遇到的种种问题，魏老师总是能够用一种严谨求实的教学态度来为我指点迷津，帮助我在继续前行。在整个论文的书写过程中，魏老师一次次为我精心修改，小到标点符号，魏老师都会帮我纠正其中的错误。在这里我想向魏老师表示我最诚挚的感谢：老师，您辛苦了！

其次，我要感谢我的研究生同学们。他们是我的舍友，是我的同门师兄弟姐妹，是我的重师的朋友们。我们都来自五湖四海，因为缘分齐聚重庆师范大学，大家在学习上相互促进，在生活上互相帮助，每当我遇到困难的时候，总是会有同学们帮助我渡过难关。在论文书写的过程中，每每遇到问题，也总是通过和同学们的交流来及时矫正自己的问题。在这里，我想向我的同学们说一声：谢谢你们！

最后，我要感谢的是我的父母。在我学习的路上，你们是我砥砺前行的坚实后盾，是我不断前进的不竭动力。正是有你们对我的大力支持和默默的付出，才让我在人生路上不断成长。

祝愿所有的你们，在以后的路上，都会一帆风顺！

邢　政

2020 年 4 月

独 创 性 声 明

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

学位论文作者签名：　　签字日期：　　年　月　日

学位论文版权使用授权书

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

学位论文作者签名：　　签字日期：　　年　月　日