

基于 VSM 的聊天机器人微信小程序设计与实现

专业：网络工程 年级：2015 级 学生：罗锡 指导教师：欧阳林艳

摘 要

在移动互联网的不断发展过程中，人工智能也随之出现，它使互联网上的信息存储数量越来越大，人们对信息需求的标准越来越高，面对目前互联网里的大量数据，快速、准确的获取数据成为当前最迫切的需求。当前人们在获取信息的时候，使用手机的用户希望能够用简便的方式实现更加快捷和更加人性化的功能，自动问答系统和聊天机器人系统的研究进入了新的发展阶段。

本文采用自动问答系统的原理和 VSM 算法进行聊天机器人的后台构建，采用微信小程序进行前台设计。聊天机器人能够采用自动问答系统进行总体设计，VSM 算法处理过程主要分为问题分析处理、算法处理、答案排序抽取三个模块。问题分析模块是将用户提出的问题进行分析；算法处理是采用 VSM 算法进行文本相似度计算，将计算结果进行排序；答案排序抽取是指将排序后的最大值进行数据库查询，查找到对应答案。采用微信小程序进行聊天机器人的前台设计，设计内容简单易懂，在设备使用上无需进行下载安装软件，只需要扫描二维码即可进行测试与使用。

通过在微信小程序上进行人机对话，能够为阳光学院的新生、家长以及业界人士带来便利，为他们提供阳光学院信息咨询服务。在设计上，以一问一答的聊天对话方式进行聊天机器人展示，满足用户使用便捷的需求；采用微信小程序来进行展示，给用户带来较好的用户体验感。在算法的选择上，采用 VSM 算法能够在检索过程中提高正确率。

关键词：自动问答系统 VSM 算法 聊天机器人 微信小程序

目 录

1 绪论	3
1.1 国内外研究现状	3
1.2 研究内容及意义	5
1.3 论文和组织结构	5
2 聊天机器人的相关技术	7
2.1 后台开发技术及工具	7
2.2 前台开发技术及工具	7
2.3 SQL SERVER 2008 R2 简介	7
2.4 分词工具	8
3 聊天机器人的构建原理	9
3.1 自动问答系统	9
3.2 向量空间模型 (VSM) 算法	10
3.3 TF-IDF 算法原理	11
3.4 算法的比较与选择	12
3.5 聊天机器人开发工具的选择	13
4 聊天机器人的需求分析	14
4.1 聊天机器人的设计目标	14
4.2 阳光学院信息需求分析	14
4.3 信息收集方式的选择	15
4.4 聊天机器人的功能需求分析	15
4.5 程序流程分析	17
4.6 数据流分析	19
5 聊天机器人的详细设计	22
5.1 聊天机器人的总体设计	22
5.2 功能模块设计	23
5.3 数据库设计	25
6 聊天机器人的实现	29
6.1 聊天机器人的自动问答实现	29
6.2 前台功能实现	30
6.3 后台功能实现	33
7 聊天机器人的测试	36
7.1 测试的评定标准	36
7.2 微信小程序界面测试	36
7.3 微信小程序与后台连接测试	37
7.4 聊天功能测试	38
结论	39
参考文献	40

1 绪论

随着互联网的不断进步，人们对信息检索需求越来越高，人们希望在检索方式上，能够获得更准确的数据信息，而非网页链接得到的数据信息^[1]。但是，在数据量如此大的互联网中，用户获取准确信息的要求成为互联网的一大难题。

然而，伴随着各种搜索引擎的出现，大数据时代的变化与发展，网络上数据信息得到了巨大的提升。但对于用户来说，虽然搜索引擎的功能变得越来越丰富，但仍然还是有很多缺点存在^[1]。一方面，搜索引擎没有全面的考虑到问题的本身存在的语义信息^[1]。在查询过程中采用关键字查找匹配方式，虽然把相关关键字信息展示出来，但是没有较好的理解句子所包含的语义信息，在句子的理解上，关键字搜索内容不能满足用户对信息的需求。另一方面，搜索引擎的结果过于庞大，其中还包含了部分广告信息以及无关信息^[1]。对于用户来说，用户在查找的过程中，需要排除广告以及无关信息，整个过程需要耗费大量时间，且不一定能够快速获得答案信息。

在用户获取信息的过程中，搜索引擎虽然很方便，但缺点也是不可避免的，它带给用户较为不好的体验，国外的相关调查公司曾经对用户搜索引擎的体验情况通过问卷调查的方式进行调查，调查数据显示，仅有百分之十八的用户认为搜索引擎能够帮助自己正确获得数据信息，而百分之六十八的用户认为搜索引擎返回的结果不能够帮助自己获得正确的数据信息，并且结果难以令人满意^[1]。所以，能够满足用户搜索需求的新型检索方式，给用户提供一种准确度更高、高效、方便、人性化的信息检索方式^[2]具有非常大的研究意义，本项目针对用户通过信息搜索，获取准确信息的需求，设计实现智能聊天机器人。

1.1 国内外研究现状

1.1.1 智能机器人发展现状

从古至今，在智能机器人技术的研究发展过程中，已经有团队花了较长时间去研究，研究过程有助于促进人机交互方式的发展^[3]。在最早的研究中，智能机器人受限于计算机数据计算能力以及当时计算机存储数据空间的大小，导致人工智能的实验规模比较狭小^[4]。经过多年的成长，通过自然语言处理、大数据分析、机器学习等等技术的发展与进步，现在的智能机器人已经实现基本聊天对话功能。

近年来，国内在智能机器人的研究上也有了较大的进步，以一些高校作为代表，其研究领域成绩比较突出，主要有清华大学、中科院计算所和哈尔滨工业大

学等^[4]。在国内的淘宝商城里，客服聊天机器人能够通过与客户聊天，使客户获得商品信息、交易信息等数据信息；国外的苹果公司研制开发的苹果手机，其自带的智能机器人 Siri^[4]，可以通过手机呼叫唤醒词，也就是其名字进行机器人唤醒，与机器人进行对话，在对话的过程中能够实现查看闹钟数据、查看天气情况、查看来电、短信信息等功能。由此可见，智能机器人的实现能够让用户在对话的过程中获取用户想要的信息，能够达到用户获取准确信息的要求。

1.1.2 自动问答系统的研究情况

在研究智能聊天机器人的过程中，如何让机器人识别并判断用户的提问数据是计算机处理过程中的一大难题，随着自动问答系统的出现，它打开了智能机器人的大门，它能够让计算机识别用户提出的问题，并且返回对应答案，也就是说，机器人能够通过计算机与人进行交流互动。

早期的自动问答系统在国外发展很迅猛，也有相应的产品出现，密歇根大学团队开发的 AnswerBus 系统^[5]主要涉及到地理、历史、文化以及日常生活等方面，还利用多种语言进行处理，技术相当成熟。然而，国内的自动问答系统发展较慢，在 1998 年才有了突破性的进展，自然语言理解模型在 1998 年被研究开发出来，并被运用到后期的应用中。现在，已经有较为成熟的自动问答系统出现，比如百度公司开发的百度知道，它充分运用到自动问答系统，它能够将问题进行自然语言处理，并将数据进行搜索、排序，将对应答案反馈给用户。

但是，在自动问答系统发展过程中，也会存在着匹配错误的情况。当用户输入的问题与数据库内所有语句进行相似度比较，结果却差强人意，这样的现象表明计算机在识别过程中依然存在着匹配错误的情况，尤其体现在对中文的文字识别情况上，因为中文在理解上比较复杂。国内对于相似度比较的研究也下了很大功夫，经过各大高校、研究院的探索与研究，中科院研究的 NLPPIR 汉语分词系统能够对中文进行分词、英文分词和关键字提取，这个分词系统对自动问答系统的中文理解有着非常重要的作用。国外对相似度的研究也做了很多贡献，比如国外学者在相似度研究过程中，提出向量空间模型^[6]（VSM:Vector Space Model），它的核心思想是将文档看作空间中的向量，并通过计算两个向量的内积来判定两个文档之间的相似度情况，内积越大代表文档相似程度越高，反之则代表相似程度越低^[6]。VSM 算法相对于其他算法而言，在工作原理的描述中，充分利用了数学公式，比较容易理解；在计算过程中，提高了正确率。

1.2 研究内容及意义

1.2.1 项目研究内容

本文将向量空间模型作为项目核心进行深度研究，采用 Java 程序开发语言进行聊天机器人核心算法的设计，JavaScript 脚本语言来设计开发一个微信小程序，通过数据抓取工具进行阳光学院数据信息的收集，对数据进行分类、整理、建模，构建知识问答数据库，最终设计出一个基于 VSM 的聊天机器人微信小程序。

1.2.2 本课题研究意义和创新点

相对于普通的聊天机器人而言，本课题的研究意义在于利用 VSM 算法进行文本相似度计算，达到提高聊天机器人回复正确率的效果。在作品的展示方面，以微信小程序的形式呈现，不需要进行下载安装软件，比普通的聊天机器人使用更方便。

本课题提出的创新点一方面是采用向量空间模型构建的机器人，较之普通的聊天机器人，问答数据能够达到更高的正确率，另一方面是针对阳光学院的数据进行收集，能够为学校提供帮助、为新生提供一些提供入学小贴士。

1.3 论文和组织结构

针对聊天机器人的构建过程，论文主要部分包括以下七个部分，最后通过结论部分进行总结，具体章节的设计部分如下：

第一部分：绪论。针对本课题的背景、国内外的现状进行了阐述，对本课题的研究内容、意义以及创新点进行分析说明。

第二部分：聊天机器人的相关技术。针对聊天机器人研发过程中所使用的技术和软件进行简单的介绍，包括后台算法开发技术、前台微信小程序的开发技术、SQL Server 2008 R2 数据库介绍以及分词工具的介绍。

第三部分：聊天机器人的构建原理。介绍了聊天机器人的构建原理，包括自动问答算法、VSM 算法的原理和 TF-IDF 算法的原理。

第四部分：聊天机器人的需求分析。通过需求分析来解析基于 VSM 的聊天机器人微信小程序项目，着重通过程序流程图、数据流图等来对项目算法进行详细的分析，阐述了对算法的选择情况、数据收集的方式。

第五部分：聊天机器人的详细设计。根据总体设计部分进行项目设计，根据流程图设计进行项目流程查看，根据功能模块分析进行主要功能模块图设计，根据逻辑结构设计、物理结构设计以及概念结构设计三个方面进行数据库设计。

第六部分：聊天机器人的实现。主要介绍基于 VSM 的聊天机器人微信小程序是如何实现的，包括 VSM 算法在聊天机器人中的应用介绍、主要的微信小程序的界面设计部分。

第七部分：聊天机器人的测试。通过测试标准对前后台进行测试，前台主要对微信小程序界面进行测试；后台主要对小程序与后台连接进行测试，聊天功能测试。最后对测试结果进行准确率分析。

结论部分。此部分主要对聊天机器人进行优缺点分析，指出不足之处，最后进行总结。

2 聊天机器人的相关技术

2.1 后台开发技术及工具

2.1.1 Java Servlet

Java Servlet 技术能够对数据库的驱动 JDBC 进行数据库的连接的实现，其对象封装了对 HTTP 请求的处理。^[7]采用 Servlet 技术能够接收处理微信小程序发送的 request 请求数据并访问数据库返回 response 回答。

2.1.2 MyEclipse 10 简介

MyEclipse 10 是一个自带 Tomcat 应用服务器，并且支持 Java Servlet 技术。Tomcat 应用服务器是一个 Web 服务器，其优点在于免费开放源代码、使用轻量^[8]。Tomcat^[8]应用服务器是一个容器，相当于中间媒介，负责处理客户发出的 HTTP 请求^[8]。MyEclipse 10 能够直接进行开发 Java Web 项目。

2.2 前台开发技术及工具

2.2.1 wx.request 网络请求方式

在微信小程序中，wx.request 能够将前台微信小程序的数据进行发送，通过一个网络请求操作发送至后台服务器，后台通过 Java Servlet 进行处理，实现前后台的交互过程。

2.2.2 微信 Web 开发者工具简介

微信小程序是一套系统的展示层（即常说的前端程序）。^[9]微信 Web 开发者工具集成了开发调试、代码编辑及程序发布等功能。^[9]通过微信 Web 开发者工具能够快速开发微信小程序，进行界面设计、JS 代码编写以及 CSS 样式设计。

2.3 SQL Server 2008 R2 简介

Microsoft SQL Server 2008 R2 是一个数据库系统，它还是由数据库、数据库管理系统、应用程序和数据库管理员组成的存储、管理、处理和维护数据的系统。^[10]它能够帮助用户进行数据存储，将问答数据存储在对数据库内。

2.4 分词工具

分词工具能够对中文、英文进行分词处理，关键字提取。NLPIR 汉语分词系统能够进行跨平台分词，以及关键字提取。因为它支持 Java 语言以及不同的开发平台。

3 聊天机器人的构建原理

3.1 自动问答系统

自动问答系统是人工智能领域和自然语言处理领域中最具有发展前景和研究方向的系统，也是聊天机器人构建的核心原理。自动问答系统能使计算机自动处理用户用自然语言提出的问题，并用准确、简洁的自然语言回答用户答案，能够使用户快速、准确地获取信息，满足用户对信息的需求^[11]。

自动问答系统的运行过程主要分为问题分析、信息检索以及答案抽取三个过程^[11]。问题分析过程是对用户提出的问题分析，信息检索过程是将用户提出的问题快速定位到数据库内，答案抽取过程是从对应数据库内查到最终答案。自动问答系统原理图如图 3-1 所示。

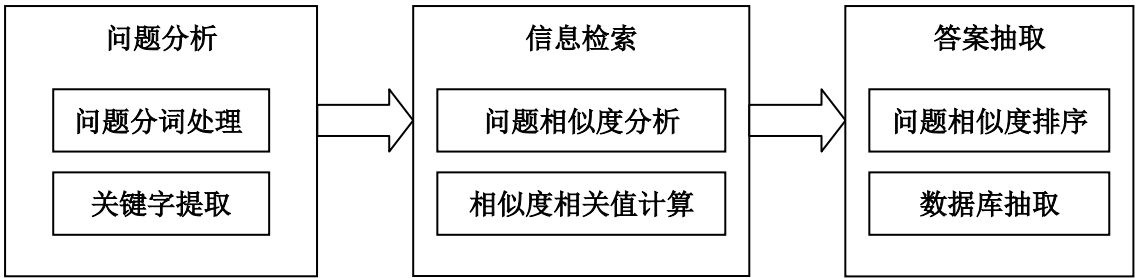


图 3-1 自动问答系统原理图

3.1.1 问题分析

问题分析是自动问答系统的第一步，是将用户提出的文字问题进行分析，从而获取用户的需求。

在问题分析过程中，主要分为两个过程。第一个过程对问题进行分词，采用分词工具对问题进行切割，为了更好的计算文本相似度。第二个过程对关键字进行提取，找到对应的数据表。

整个问题分析过程中，比如用户提出：“从福州南站怎么去阳光学院”，经过分词工具，得到对应的分词结果是“从/福州/南站/怎么/去/阳光学院”，关键字提取情况为“福州”、“南站”、“阳光学院”。通过关键字提取，从而知道用户想要问的问题，也就说明，通过对问题进行分析能够得知用户想要获取阳光学院的路线信息。此时就能够快速定位到路线表中，提高了检索速率。

3.1.2 信息检索

信息检索是自动问答系统的第二步,是将问题分析处理后的数据进行快速定位,查找到计算机的数据库内最相近的问题。

在信息检索过程中,将问题分析后的问题以及关键字定位到的数据表信息一起传递到信息检索中,采用 VSM 算法和 TF-IDF 算法进行信息检索,将数据表对应数据临时存入 doc 中,统计对应的 TF-IDF 值、词频等数据,以便后期进行分析。

3.1.3 答案抽取

答案抽取是自动问答系统的第三步,是将信息检索后的问题进行数据库查找,找到问题回答。

在答案抽取过程中,聊天机器人主要将信息检索后的值用键值对进行存储,用 compare 函数进行排序,将最大的键值对输出,利用最大值的 ID 进行数据库查询,寻找与用户提出问题最接近的答案。

3.2 向量空间模型 (VSM) 算法

向量空间模型 (VSM) 是近年来较为简单的一个文本相似度算法。通过将一篇较为完整的文章看作文档 $D(T_1, T_2, \dots, T_n)$,将文档内的每一句话视为词条 T_i ,对于含有 n 个项的文档 $D(T_1, T_2, \dots, T_n)$ 都会有一个代表重要程度的权重 W_k ,且满足 $D=D(W_1, W_2, \dots, W_n)^{[12]}$ 。

在 VSM 中,文本指的是用汉字书写的内容,可以是词语,也可以是句子;特征量,也就是词条,它构成的集合就是文本内容;特征量的权重是代表重要程度的权重,不同的词条对应一个不同的权重,权重的大小在相似度值计算的过程中很重要^[12]。

VSM 算法的计算原理是把文档 D 中的每一个词条 $T_1, T_2, \dots, T_n$ 看成一个坐标系,文档看成 A 和 B ,利用数学知识进行计算余弦值。余弦夹角值^[13],也就是相似度值 $\text{Sim}(D_1, D_2)$,可以用来衡量文档 D_1 和 D_2 之间的相关程度。其对应余弦夹角图如图 3-2 所示。

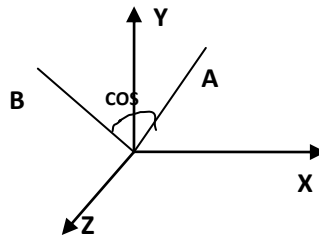


图 3-2 余弦夹角图

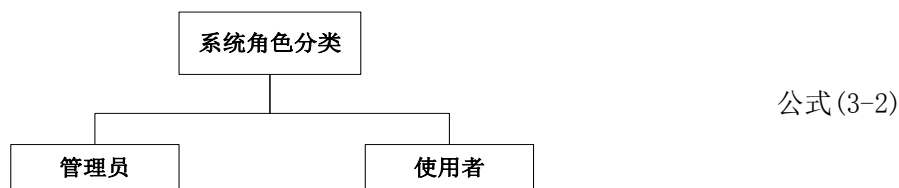
用向量矩阵的方式表示文本 D。^[14] 向量矩阵图如图 3-3 所示。

$$D = \begin{bmatrix} D_1 \\ D_2 \\ \dots \\ D_N \end{bmatrix} = \begin{bmatrix} W_{11} & W_{12} \dots W_{1K} \\ W_{21} & W_{22} \dots W_{2K} \\ \dots & \dots \\ W_{N1} & W_{N2} \dots W_{NK} \end{bmatrix}$$

图 3-3 向量矩阵图

根据图得出对应公式 3-1 所示，也可以用公式 3-2 来表示。

$$Sim(D_1, D_2) = \sum_{k=1}^n W_{1k} * W_{2k} \quad \text{公式 (3-1)}$$



公式 (3-2)

其中， W_{1k} 、 W_{2k} 分别表示文档 D_1 和 D_2 中的权值， $1 \leq k \leq N$ 。

3.3 TF-IDF 算法原理

TF-IDF 算法是对词语出现的频率进行统计与计算，TF（即词频）是指在文档中某一词语，给词语指定不一样的次数来进行计算，对于出现频率比较高的词

语，比如“你”，“我”，“的”，应该将这种类型的词频设置一个较低的值，使之不能影响到计算结果。对于某些出现频率小，却能说明文档关键字的词语，应该将这类的词频设置较高的值，这类的词语出现的频率称之为 IDF（即逆向文件频率），它是 TF 值的倒数^[15]。

TF 的计算等于词语出现的次数除以该文档的总词语数。而 IDF 值的计算等于总文档数目除以包含该词语的文档数目，再将算出的商求出对数值，也就是 $IDF = \log(N/n_k)$ ^[15]。在 TF-IDF 算法中权重 w 是根据 TF 与 IDF 的值计算得来的，通过整理得出公式 3-3 所示。

$$w_{ik} = \frac{tf_{ik} * \log\left(\frac{N}{n_k}\right)}{\sqrt{\sum_{k=1}^n \left(tf_{ik} * \log\left(\frac{N}{n_k}\right)\right)^2}} \quad \text{公式 (3-3)}$$

3.4 算法的比较与选择

在聊天机器人的设计中，有两种方案，方案一采用空间向量模型直接将文本中出现的词语作为词条进行相似度计算；方案二将向量空间模型与 TF-IDF 算法综合进行计算。

采用方案一进行计算相似度计算过程中，也就是采用公式 3-2 进行计算，根据计算结果显示，在文档 M 的文本 1 中有 100 个词语，词语 A 出现的频率为 5 次，文本 2 中有词语 50 个，同样的词语 A 出现的频率却是 4 次，在计算相似度时，前一个的结果要远大于后一个的结果，显然这是不合理的，因为前一个词语占文本的 5%，后一个为 8%。

采用方案二进行相似度计算时，也就是采用公式 3-3 进行计算，TF-IDF 算法能够在计算过程中对词语出现的频率进行统计与计算，并将“你”、“我”、“他”这样的词语设置较小的频率，突出关键字的频率，提高了算法的相似度计算，不会出现方案一的问题。

在单独采用空间向量模型进行相似度计算时，也会遇到一些问题，情况一是当文件内容太长，会导致内积变小，算出的相似度值不准确，不适用与数据较长的文本；情况二，没有对词频数据进行统计，结果会产生一定的误差^[16]。

综合来说，采用方案一在中文汉字的文本相似度计算时会出现较大误差，方案二能够提高向量空间模型的算法计算能力，所以采用方案二。

3.5 聊天机器人开发工具的选择

聊天机器人的开发工具选择决定了最后聊天机器人开发模式，所以它是相当重要的。

后台开发工具有三个软件，分别是 Eclipse、NetBeans 和 MyEclipse。NetBeans 主要适用于界面设计与开发，当前不适合后台算法环境的搭建。Eclipse 可以进行后台开发，但需要下载安装 Tomcat 软件才能进行接收 HTTP 请求，需要花一些时间进行环境搭建。MyEclipse 自带有 Tomcat 软件，可以直接开发 Java Web 项目，适用于当前开发。综合来看，后台采用 MyEclipse 作为开发工具。

前台展示方式有很多，但为了用户使用方便，所以采用微信 web 开发者工具进行微信小程序的开发，微信小程序使用简单易懂，便于用户操作。

4 聊天机器人的需求分析

4.1 聊天机器人的设计目标

通过官网来查看学校信息，只能粗略了解到学校基本信息，并不能满足新生对大学信息的需求，为了让新生更全面的了解到路线、食堂等等学校信息，设计了基于 VSM 的聊天机器人微信小程序，通过小程序的使用，能够让新生得到一些入学小贴士，满足他们的需求。

聊天机器人分为两部分，一部分是前台进行人机交互以及提出反馈情况，第二个部分是后台采用 VSM 算法进行数据处理。前台的人机交互部分指的是用户通过小程序提出问题进行与聊天机器人的连接，达到与后台的交互；提出反馈情况部分指用户对问答数据提出反馈。后台数据处理部分指的是对用户提出的问题进行分词、VSM 算法等等处理，最后通过在问答知识库内查找并获得问题的答案。通过这两部分的设计，实现一个能够帮助新生、家长以及业界人士更好的了解阳光学院的微信小程序。

4.2 阳光学院信息需求分析

针对大学学生、家长以及业界人士对阳光学院信息需求的调查，得出以下几种情况。

第一类是大学生，对于即将进入大学的新生，其主要信息需求分为学校大致情况、专业情况、师资情况、宿舍情况、食堂情况、路线情况。

第二类家长，对于家长来说，其主要信息需求分为学校大致情况、专业情况、师资情况、就业情况、入学小贴士、宿舍情况、食堂情况。

第三类业界人士，对于业界人士来说，其主要信息需求分为学校大致情况、专业学习情况、发展情况、就业情况。

针对以上情况，对数据进行综合整理，结合智能聊天机器人自身数据，将智能聊天机器人的所需要的信息归纳如下：

- (1) 智能机器人本身基本信息；
- (2) 学校基本情况，包括食堂、宿舍情况；
- (3) 信息工程学院专业情况，包括专业师资情况、就业情况、发展情况；
- (4) 入学小贴士；
- (5) 路线情况。

4.3 信息收集方式的选择

对于智能机器人本身基本信息、阳光学校基本情况，主要对学校食堂、学校宿舍的情况进行数据收集、阳光学院信息工程学院大致情况，包括专业师资情况、就业情况、发展情况、入学小贴士、路线情况这些数据，在收集的过程中，如果只采用人工收集，那么需要花大量的时间进行收集。所以，在信息收集方式的选择上，主要采用人工收集和相关数据抓取软件进行收集，比如八爪鱼。具体分配情况如下：

（1）人工收集数据，主要收集智能机器人本身基本信息、食堂、宿舍情况、入学小贴士、路线情况。

（2）数据抓取软件收集数据，主要在百度知道、贴吧等网站收集学校基本情况、信息工程学院专业情况、专业师资情况、就业情况、发展情况。

4.4 聊天机器人的功能需求分析

基于 VSM 的聊天机器人微信小程序设计过程中，用户无需注册，通过扫描二维码在微信小程序内进行提问，提出你感兴趣的问题，机器人根据 VSM 算法得出最靠近的问答库内的问题，根据问题 ID 查找对应答案，返回给用户。小程序整个部分分为微信小程序模块和后台核心 VSM 算法模块。

微信小程序模块包括提问模块和反馈模块。

（1）提问模块：用户在微信小程序内提出感兴趣的问题，并能够在提问页面查看到答案回复情况。

（2）反馈模块：用户在微信小程序内反馈页面提出反馈，查看反馈情况，根据反馈来查看小程序的正确率。

微信小程序模块情况图如图 4-1 所示。

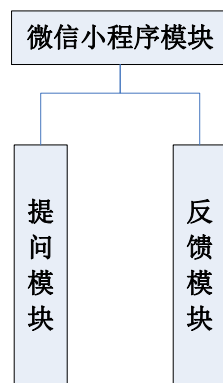


图 4-1 微信小程序模块情况图

后台核心 VSM 算法模块包括查询答案模块、问题检索模块、文字预处理模块、用户交互数据处理模块。

(1) 用户交互数据处理模块：包括数据接收和数据返回。数据接收指在后台对微信小程序发来的数据进行接收，转化为对应格式。数据返回指将答案返回到微信小程序内。

(2) 文字预处理模块：包括分词处理和关键字提取。利用分词工具将转化后的数据进行处理，并进行关键字提取，获得问题的核心。

(3) 问题检索模块：包括相似度比较和相关值排序。将分词后的问题与数据库内数据进行相似度比较，找到最大的 Sim 值。

(4) 答案查询模块：包括数据库匹配和答案查找。通过将最大的 Sim 值对应的 ID 进行数据库查询匹配，找到对应答案。

后台核心 VSM 算法模块情况图如图 4-2 所示。

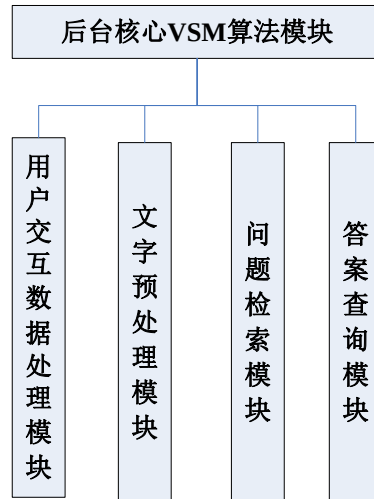


图 4-2 后台核心 VSM 算法模块情况图

4.5 程序流程分析

通过程序流程图可以看到该项目的总体流程情况，包括微信小程序与后台 VSM 算法交互的过程，程序流程图反应了程序的具体实现过程。前后台交互实现程序流程图如图 4-3 所示。

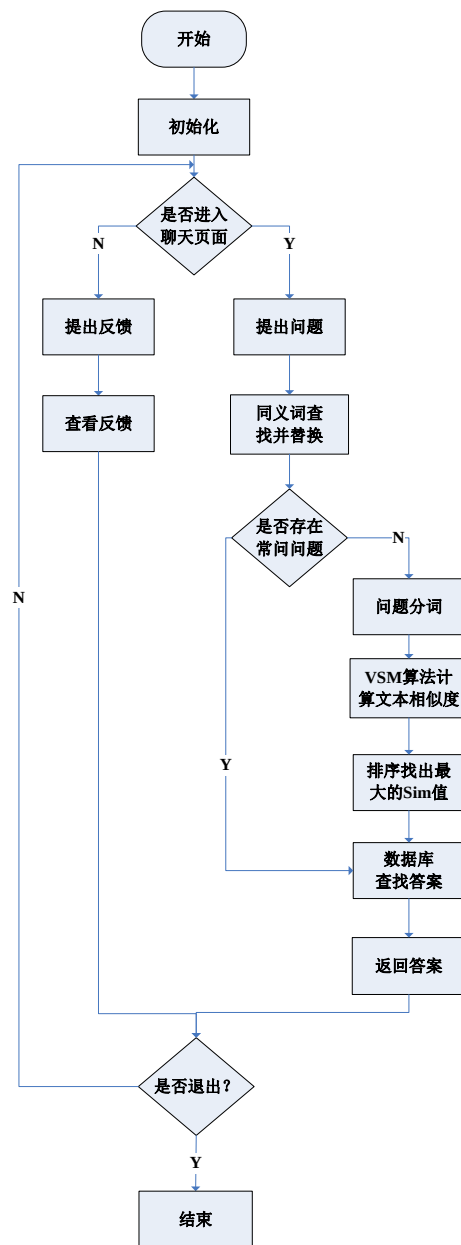


图 4-3 前后台交互实现程序流程图

根据项目的总体流程，通过时序图区分用户与核心算法处理之间的联系。用户自动问答时序图如图 4-4 所示。

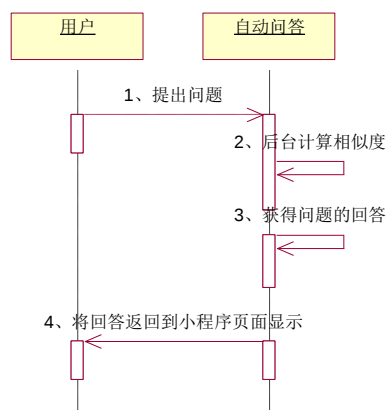


图 4-4 用户自动问答时序图

核心算法处理即后台 VSM 聊天机器人，其对应的运行时序图如图 4-5 所示。

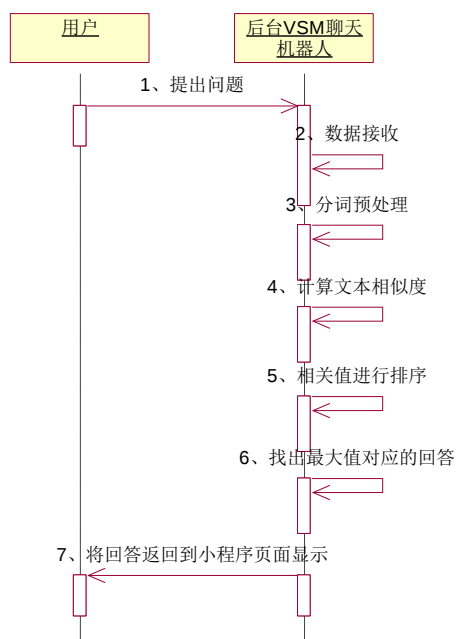


图 4-5 后台 VSM 聊天机器人的运行时序图

4.6 数据流分析

通过数据流图（DFD）分析，可以分析出项目完成的逻辑功能，DFD 以图形的方式描述数据流从输入部分到输出部分的变化过程。

4.6.1 顶层数据流图

顶层数据流图，反映了项目的主要经过，该项目描述了微信小程序与后台之间的联系。在基于 VSM 的聊天机器人微信小程序中，对数据进行加工处理的是基于 VSM 的聊天机器人。具体的顶层数据流图如下图 4-6 所示。

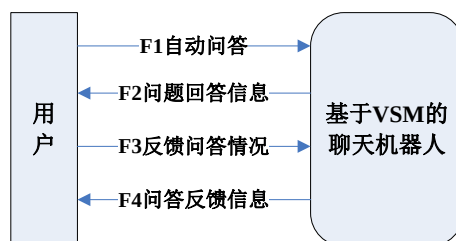


图 4-6 顶层数据流图

4.6.2 第 0 层数据流图

0 层数据流图是对顶层数据流图的扩充，将加工即基于 VSM 的聊天机器人进行扩充，得出加工项包括自动问答和反馈情况管理。第 0 层数据流图如图 4-7 所示。

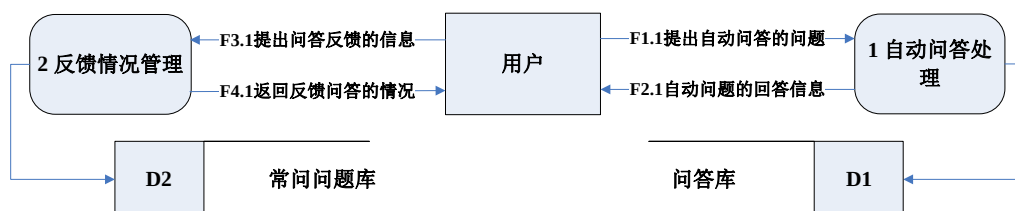


图 4-7 第 0 层数据流图

4.6.3 第 1 层数据流图

将自动问答系统处理进行细化，得出加工为数据处理、文字预处理、问题检索、答案抽取。第 1 层数据流图如图 4-8 所示。

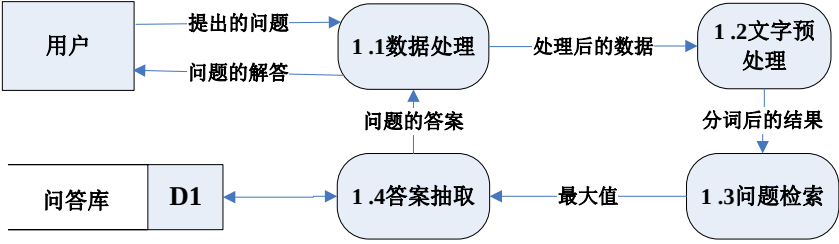


图 4-8 第 1 层数据流图

5 聊天机器人的详细设计

5.1 聊天机器人的总体设计

聊天机器人总体设计原理采用自动问答系统，在项目里自动问答系统是一个用计算机处理用户提出的问题，并将数据库内查询到的答案用准确、简洁的自然语言回答给用户的系统。它能够使用户快速、准确地获取信息，满足用户对信息的需求。

自动问答系统的原理主要体现在问题分析、信息检索和答案抽取三个过程中。问题分析过程是对用户提出的问题进行分析，利用中科院分词工具 NLPTR 对问题进行分词，提取关键字，通过关键字的分析得出问题核心。信息检索过程是将分词后的问题利用 VSM 算法得出最大的 Sim 值，用 Sim 值找出最接近数据库内的问题，获得快速定位到对应数据库中。答案抽取过程是根据最接近的问题对应的 ID 来查找最终答案。

5.1.1 总体设计框架

聊天机器人总体设计框架如下图 5-1 所示。

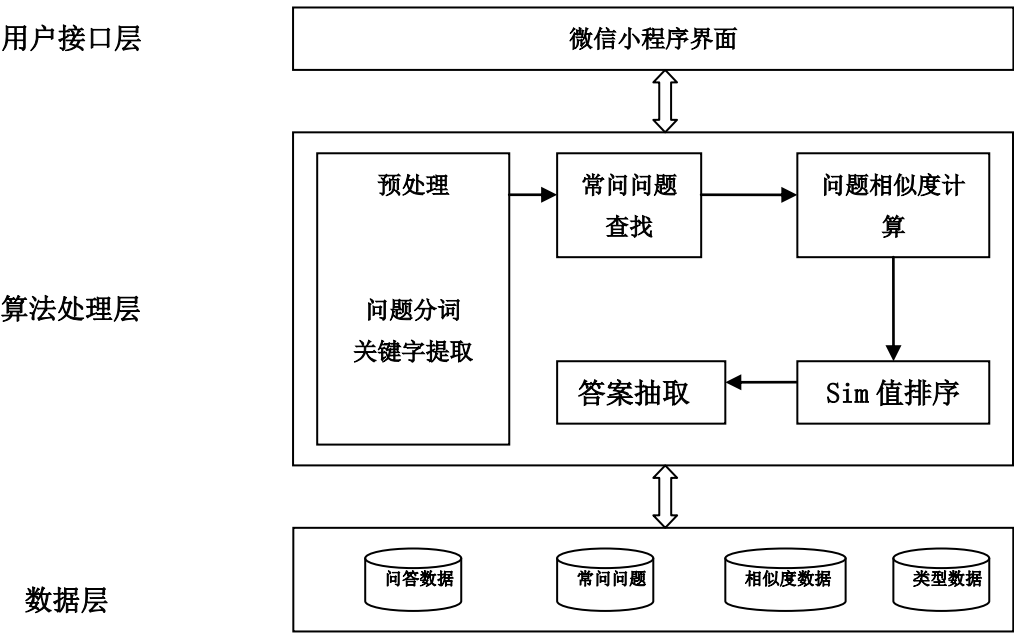


图 5-1 聊天机器人总体设计框架

从用户的角度分析，在微信小程序界面与后台进行交互的过程，得出用户的用例图如图 5-2 所示。

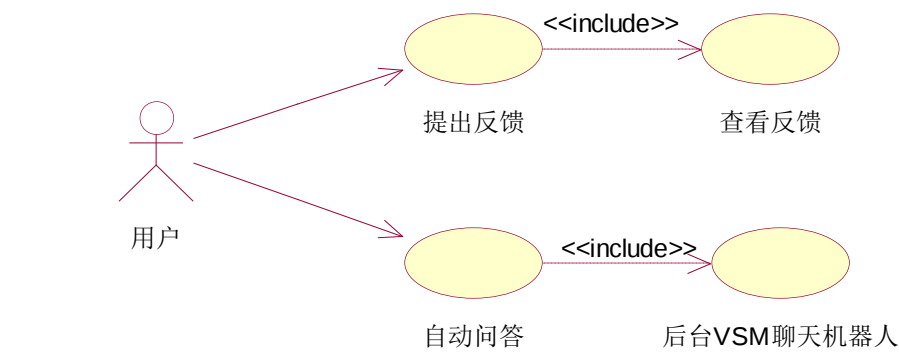


图 5-2 用户用例图

后台 VSM 聊天机器人，是一个对用户提出问题的处理过程，对算法处理层的后台用例图进行分析，如图 5-3 所示。

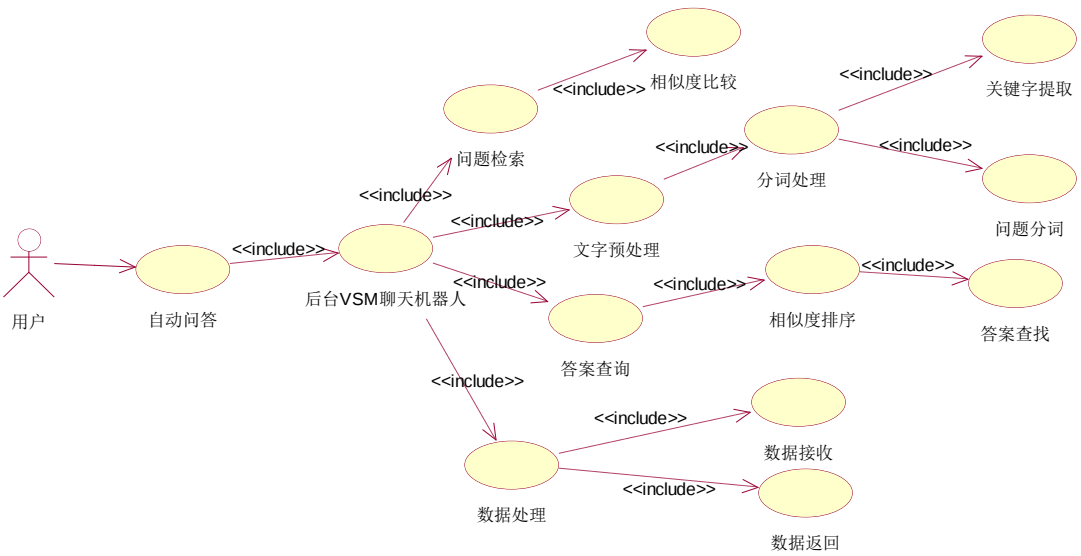


图 5-3 后台 VSM 聊天机器人用例图

5.2 功能模块设计

采用功能模块图对项目进行功能模块设计，能够清晰的表达出项目的主要功能。

5.2.1 聊天机器人微信小程序的主要功能

根据需求分析，并结合向量空间模型（VSM）算法技术，对整个框架进行综合设计，在主要功能模块图中，将以上的后台 VSM 聊天机器人细化处理，具体分为用户交互数据处理模块、问题检索模块、文字预处理模块、答案查询模块。通过细致的描述能够较为清晰的看出基于 VSM 的聊天机器人微信小程序的主要功能模块。聊天机器人功能结构图如图 5-4 所示。

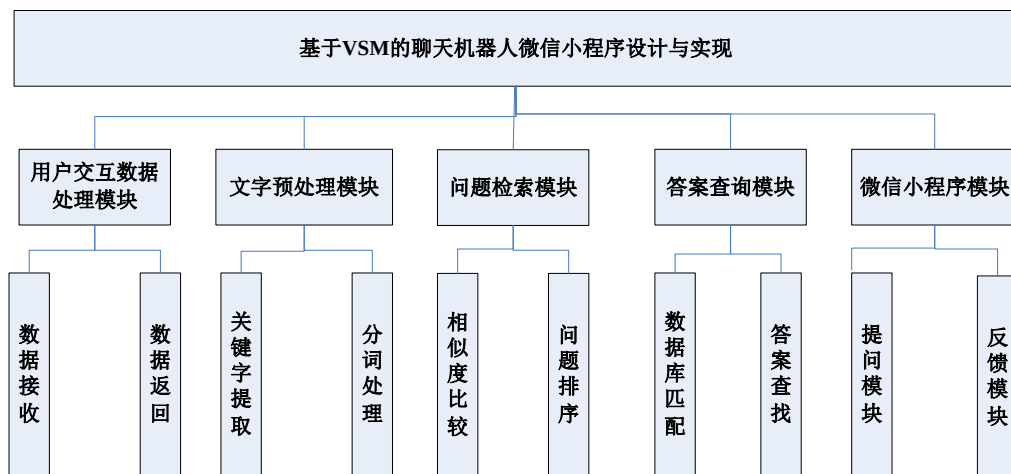


图 5-4 聊天机器人功能结构图

用户与后台 VSM 的聊天机器人的交互活动图能够看到用户提问后，后台处理的过程。交互活动图如图 5-5 所示。

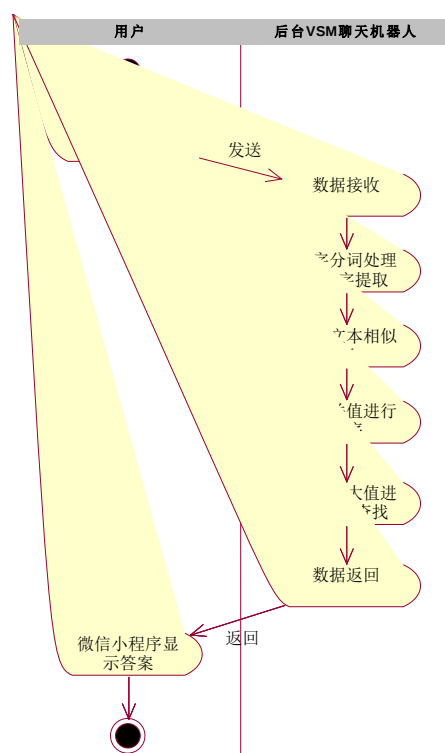


图 5-5 交互活动图

5.3 数据库设计

综合上述各类用户的信息需求，将信息需求转换为用户问答数据，数据具体分为以下几种类型。

第一种为问答数据，主要是关于机器人的数据表、关于学校的数据表、关于路线的数据表、关于信息工程学院的数据表，这类表的主要内容包括问题 ID、问题、答案、类型 ID、分词后的问题、tf、idf、w 值、排序后的 w 值。

第二种为类型数据表，表内应有类型 ID、类型关键字，可以根据类型进行库内查询，较快定位到数据表内。

第三种为常问问题表，表内应有问题 ID、答案、提出的问题、提出问题的次数，通过问题 ID 可以快速查找到答案，减少索引时间。

第四种为 TF-IDF 数据表，包括词语、词语出现的频率、TF-IDF 相关数据。

5.3.1 概念结构设计

(1) 根据功能需求分析，找出实体

实体：机器人数据、学校数据、路线数据、学院数据、类型、常问问题、TF-IDF 数据。

(2) 实体间的联系

机器人数据、学校数据、路线数据、学院数据统称为问答数据。

类型与问答数据的属于联系是一对多；TF-IDF 数据与问答数据的计算联系是一对一。常问问题与问答数据的统计联系是一对一。

根据以上分析，得出项目的 E-R 图，如图 5-6 所示。

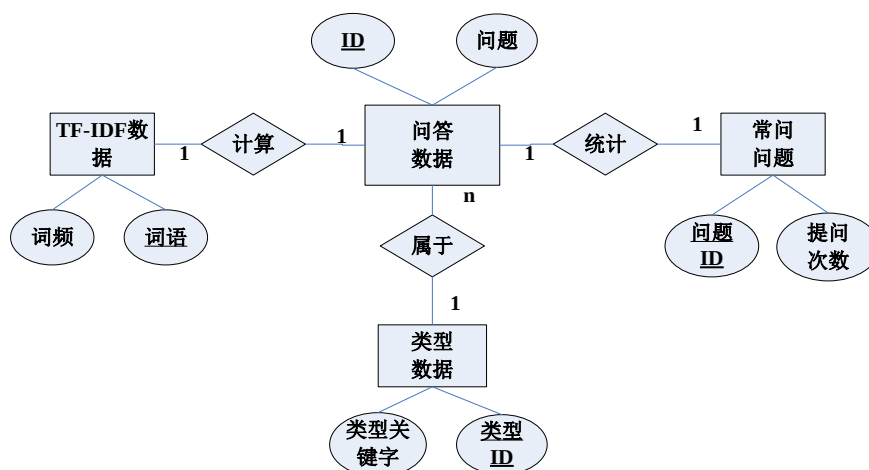


图 5-6 E-R 图

问答数据包括机器人数据、学校数据、路线数据、学院数据，对应属性相同，用问答数据来表示。图中有部分属性被省略，详细的说明如下：

问答数据：问题 ID、问题、答案、类型 ID、分词后的问题、w 值。

类型数据：类型 ID、类型关键字。

常问问题：问题 ID、答案提出的问题、提出问题的次数、答案。

TF-IDF 数据：词语出现的频率，词语，TF 值，IDF 值。

5.3.2 逻辑结构设计

根据 E-R 模型，转化得出关系模式。

(1) 实体转换为关系模式

问答数据（问题 ID、问题、答案、类型 ID、分词后的问题、w 值）为机器人数据、学校数据、路线数据、学院数据实体对应的关系模式，其中，问题 ID 是主键，类型 ID 是类型数据的外键。

类型数据（类型 ID、类型关键字）是类型数据的关系模式，其中，类型 ID 是主键。

常问问题（问题 ID、答案提出的问题、提出问题的次数、答案）是常问问题的关系模式，其中，问题 ID 既是主键也是问答数据的外键。

TF-IDF 数据（词语、词语出现的频率、tf 值、idf 值）是 TF-IDF 数据的关系模式。其中，词语是主键。

(2) 联系转化为关系模式

在概念结构设计阶段，有三个联系，一个是问答数据实体和计算联系之间的关系，它是一个 1 对 1 的联系，可以将其放在问答数据实体转化为的关系模式上。另一个是问答数据实体和统计联系之间的关系，它是一个 1 对 1 的联系，也可以将其放在问答数据实体转化为的关系模式上。最后一个是问答数据实体和属于联系之间的关系，它是一个 n 对 1 的联系，也可以将其放在问答数据实体转化为的关系模式上。

5.3.3 物理结构设计

本项目对每个表结构进行设计，设计情况如下。
robot 机器人数据表、university 学校数据表、route 路线数据表、academy 学院数据表结构一致，如表 5-1 data 数据表所示。

表 5-1 data 数据表

字段名	数据类型	主外键	是否为空	字段说明
ID	int	主键	否	问题 ID
question	varchar(20)		否	问题

续表 5-1

字段名	数据类型	主外键	是否为空	字段说明
NLPIR_question	varchar(20)		否	分词后的问题
answer	varchar(50)		否	答案
w	varchar(20)		否	w 值
type_id	int	外键 (Type)	否	类型 ID

通过类型表，可以快速定位到对应数据表中。type 类型表结构如下表 5-2 所示。

表 5-2 Tpye 类型表

字段名	数据类型	主外键	是否为空	字段说明
-----	------	-----	------	------

type_id	int	主键	否	类型 ID
type_key	varchar(20)		否	类型关键字

通过常问问题表的设置，能够让用户在问题搜索时减少检索时间，如表 5-3 所示。

表 5-3 often_ask 常问问题表

字段名	数据类型	主外键	是否为空	字段说明
ID	int	外键(data)、主键	否	问题 ID
question	varchar(20)		否	问题
submit_time	int		否	提出次数
answer	varchar(20)		否	答案

TF-IDF 表储存词语相关数据以及 TF-IDF 数据。TF-IDF 数据表结构如下表 5-4 所示。

表 5-4 TF-IDF 数据表

字段名	数据类型	主外键	是否为空	字段说明
word	varchar(10)	主键	否	词语
word_time	varchar(20)		否	词语出现频率
tf	varchar(20)		否	Tf 值
idf	varchar(20)		否	Idf 值

6 聊天机器人的实现

该项目在基于 VSM 的聊天机器人微信小程序的实现过程中,采用前后台开发模式进行项目开发,采用 SQL Server 2008 R2 数据库进行问答数据的存储。在该项目中,数据库的静态表较多,表结构与表内数据也相对较为全面。

在前台开发过程中,微信小程序采用 JavaScript 语言开发聊天机器人,在 JavaScript 中采用 `wx.request` 方式进行数据传输,在微信 web 开发者工具中进行界面设计。

在后台开发过程中,为了较好的处理微信小程序提出的问答数据请求,主要用 Java Web 进行开发、MyEclipse 作为开发工具,采用 Servlet 技术处理问答数据请求并作出响应,将相关数据回复到微信小程序显示。在项目中,主要采用 NLPIR 的中文分词和对问题的关键词提取两大功能,进行问题的初步处理,也就是对用户提出的问题进行分词、关键字提取,问题的相似度处理采用 VSM 算法以及 TF-IDF 算法。

6.1 聊天机器人的自动问答实现

首先,用户通过扫描二维码登录微信小程序,在微信小程序的首页查看到阳光学院的简介,通过点击跳转至聊天页面,把你想要发送的数据通过按钮发送到指定网址,也就是后台。

其次,后台获取到微信小程序发送的数据,将数据进行解析,获得用户提出的问题,将问题进行分词处理,得到分词结果和关键字,通过关键字查找到对应数据库。此时,调用 VSM 算法函数,将数据库信息和分词结果传递过去,通过计算获得 TF 值、IDF 值以及 Sim 值等数据,将这些数据进行存储。将所有的 Sim 值和对应问题进行排序,找到最大值对应的问题,利用问题对应的 ID 进行数据库查找,从对应数据库中查到答案。

最后,将答案返回到微信小程序显示,让用户得到问题的解答。这就是该项目自动问答的实现全过程。

聊天页面的聊天情况图如 6-1 所示。

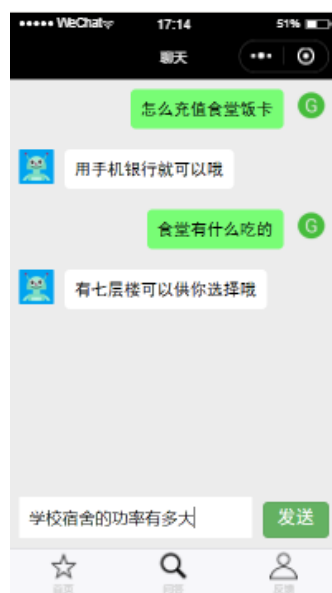


图 6-1 聊天情况图

反馈页面的反馈情况图如 6-2 所示。

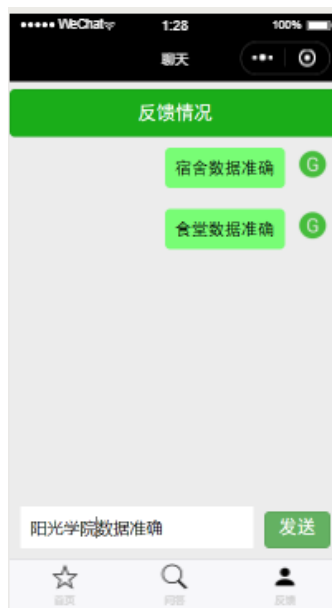


图 6-2 反馈情况图

6.2 前台功能实现

前台功能实现主要体现在前台微信小程序实现数据传输、聊天等界面设计、

反馈设计。

6.2.1 实现数据传输

在微信小程序里，用户需要通过小程序聊天页面发送问题，所以，实现问题数据的传输是项目实现的第一步，也是最为关键的一步。为了问答数据能够完整的发送至后台，在项目中，主要采用 JavaScript 里的 request 方法进行数据传输。具体 js 代码如下。

```
wx.request({  
  url:"http://192.168.1.100:8080/Robot/getServlet",header:  
{ "Content-type": "application/x-www-form-urlencoded;charset=UTF-8"},  
  data:{info:msg.content},  
  method:'POST',  
})
```

6.2.2 实现微信小程序页面制作

在基于 VSM 的聊天机器人微信小程序过程的实现中，微信小程序需要实现首页、聊天页和反馈页三个页面。

首页对基于 VSM 的聊天机器人微信小程序的实现部分进行简要介绍，用户可以通过首页了解相关信息，并在此页面通过点击问答能够跳转到聊天页面进入聊天。首页图如 6-3 所示。



图 6-3 首页图

聊天页是基于 VSM 的聊天机器人微信小程序实现的最主要的页面，用户可在聊天页进行提问，获得问题的解答，在此页面也可通过点击反馈跳转至反馈页，说明反馈情况。聊天页如图 6-4 所示。

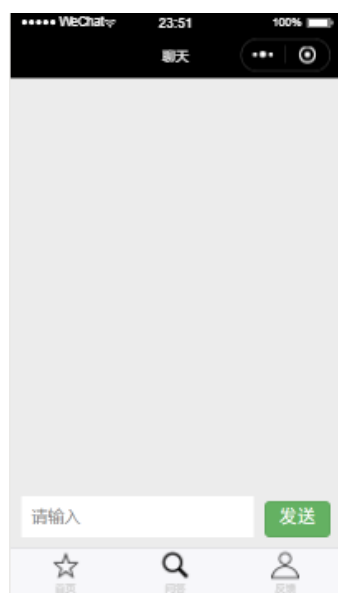


图 6-4 聊天页

反馈页能够提出反馈，通过反馈情况能够获得相关数据准确与否的信息，能够在反馈情况看到提出的反馈信息。以下是对反馈页的设计。如图 6-5 所示。

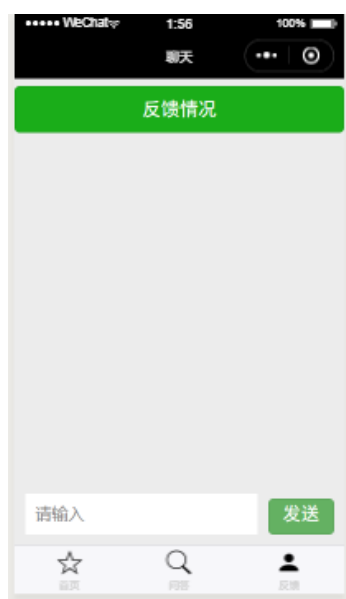


图 6-5 反馈页

6.3 后台功能实现

后台功能实现主要体现在实现算法的问答数据处理过程。这个功能是聊天机器人的核心部分。

6.3.1 实现问答数据处理

基于向量空间模型（VSM）的问答数据处理主要包括分词工具的使用、向量空间模型以及数据库查询。

分词工具采用中科院分词工具 NLPIR，对问题进行分词处理以及关键字提取，对提取的关键字进行模糊查询，找到对应的表。

向量空间模型，即 VSM 算法，它的工作原理是将 TXT 文档进行识别并实现临时存储，然后进行计算 TF 值、IDF 值以及 Sim 值等数据。在该项目中，首先将用户提出的问题与数据库问题数据进行识别并实现临时存储，以便后续的计算。其次，利用 VSM 计算值的方法进行相关数据的计算，将数据进行存储，以便查看。最后，将所有的 Sim 值和对应问题进行排序，找到最大值对应的问题，利用问题对应的 ID 进行数据库查找，从对应数据库中查到答案，返回给用户。

（1）根据 TF-IDF 值的计算过程，代码如下所示。

```
int i = 0;
if(i < size){ //i 小于文档词数的长度
```

```

s = doc[i]; //将分词处理后的数据存入 s 中
if(tf.containsKey(s)){ //判断 tf 值是否存在
    tf.put(s, tf.get(s)+1); //读取 tf 数据到数据集
}
else {
    tf.put(s, (float) 1.0);
    if(idf.containsKey(s)) //判断 idf 值是否存在
        idf.put(s, idf.get(s)+1); //读取 idf 数据到数据集
    else
        idf.put(s, (int) 1);
}
++i;
}

```

(2) 对向量的模进行处理，计算 Sim 值 (cos 值) 的计算过程如下所示。

```

int i = 0;
if(i < len){ //i 小于总文档数
    tf2=tfidf[i].keySet(); //获取到键值对里 key 值，tfidf 值等于
    tf*idf
    for(String str:tf2){
        x[i] += tfidf[i].get(str)*tfidf[i].get(str); // 进 行
        tfidf*tfidf 计算
    }
    x[i] = (float) Math.sqrt((float) x[i]); //对 idf*idf 进行开根号
    处理
    ++i;
}
ans += tfidf[x].get(s)*tfidf[y].get(s); //进行 tfidf*tfidf 计算
every_sim=ans/(x[x]*x[y]); //计算每个句子的 Sim 值

```

(3) 查找最大的 Sim 值代码主要通过 sort 方法进行键值对的排序，用 compare 方法进行键值对对比，最后通过 getKey 将最大的 key 值查找处理。代码如下所示。

```

Collections.sort(data, new Comparator< data.Entry<String, String>>()
{
    //利用 Collections 里的 sort 方法对 data 进行排序

```

```
public int compare(data.Entry<String, String> o1,  
data.Entry<String, String> o2) { //将参数 o1 和 o2 进行比较  
    return (o2.getValue()).compareTo(o1.getValue()); // 利用  
compareTo 比较器，对 o1 和 o2 进行降序排序  
}  
});  
Object maxId = data.get(0).getKey(); //获取最大值对应的 key 值
```

7 聊天机器人的测试

本章对该项目进行测试，主要包括微信小程序界面测试、微信小程序与后台连接测试和聊天功能测试，并对结果进行分析。

7.1 测试的评定标准

测试的评定标准对聊天机器人的测试非常重要，项目后期可以通过评定标准来判断问答数据的准确性，并且进行后期修改。所以，对于测试结果的评定标准，主要有以下两个模块进行判断。

方法一是以阈值（即最大值）进行判断。阈值的确定进行项目评定的标准之一，当阈值小于 0.5，问答数据为不准确的回答。当阈值大于 0.5 时，问答数据都与数据库内数据有关联，能够作为回答；当小于 0.5 时，问答数据偏离数据库内数据，不足以作为回答。采用 VSM 算法对问答数据进行相似度比较，将 Sim 值进行排序，如果得到的最大值小于 0.5，则返回固定回答，此条数据视为不准确回答。

方法二通过人工判断。在聊天功能的测试上，用户是项目的体验者，他们的体验感受能够为项目提出建议，帮助项目后期的整改。所以，人工判断的形式必不可少。

7.2 微信小程序界面测试

微信小程序界面测试通过微信小程序的界面情况进行测试，验证项目是否符合要求。

通过相关用例测试对微信小程序的界面进行测试，如下表 7-1 所示是对微信小程序各个页面跳转按钮能够跳转至对应位置进行测试。

表 7-1 页面跳转按钮测试

界面描述	在微信小程序里，点击不同按钮达到跳转到不同页面的效果
用例目的	点击按钮能否实现正确跳转
前提条件	各页面已设计完毕，能够正常运行
测试内容	点击按钮聊天是否跳转到聊天页面；点击按钮反馈是否跳转到反馈页面；点击跳转返回是否跳转到首页；

预期结果	按钮能够实现正确跳转
测试结果	按钮能够实现正确跳转

测试后的结果与预期结果相符合，测试结果表明该项目能够及时响应按钮操作。

如下表 7-2 所示是对微信小程序各个页面实用度进行测试。

表 7-2 页面实用度测试

界面描述	通过微信小程序里每个页面位置摆放合理
用例目的	页面摆放位置是否合理，是否有位置太小、无法点击的情况
前提条件	各页面已设计完毕，能够正常运行
测试内容	在首页、聊天页、反馈页能否较快找到对应按钮，在页面设计上，是否合理、是否出现位置太小、无法点击的情况
预期结果	每个页面能够较快找到对应按钮，页面设计较为合理、没有出现位置太小、无法点击的情况
测试结果	每个页面能够较快找到对应按钮，页面设计较为合理、没有出现位置太小、无法点击的情况

测试后的结果与预期结果相符合，测试结果表明该页面设置能够基本符合大部分人的需求。

7.3 微信小程序与后台连接测试

微信小程序与后台连接测试主要通过微信小程序发送数据，后台能否接到数据，数据接收情况进行测试，验证该项目是否符合要求。

如下表 7-3 所示是对微信小程序发送数据，后台能否正确接收数据进行测试。

表 7-3 数据接收测试

数据描述	微信小程序发送的数据，后台能够接收到且数据正常，无乱码
用例目的	后台能否收到小程序发送的数据，数据是否出现问号乱码情况
前提条件	前后台已搭建完毕，能够正常运行
测试内容	微信小程序发送 20 次数据至后台，在后台观察数据是否正常接收，是否有出现问号、乱码现象

预期结果	失误率不超过 0%，未出现问号、乱码情况
测试结果	失误率不超过 0%，未出现问号、乱码情况

测试后的结果与预期结果相符合，测试结果表明后台能够成功接收每一次数据，而且后台接收的数据不会出现乱码、问号的情况。

7.4 聊天功能测试

聊天功能测试主要对聊天数据进行测试，并对准确率进行分析。

在测试过程中，对聊天功能的测试一方面主要看阈值的大小，是否小于 0.5，如果小于 0.5，那么表示该条数据回答不够准确，返回固定回答。另一方面看用户反馈情况。通过用户的反馈数据能够判断数据的准确率。在本文中，由于用户较少，所以主要针对阈值数据进行测试。

如下表 7-4 所示是聊天准确率的测试情况。

表 7-4 聊天准确率测试

问题分类	提问次数	得到正确答案次数	准确率
对机器人的提问	50	35	0.70
对学校信息的提问	50	36	0.72
对食堂信息的提问	50	40	0.80
对路线信息的提问	50	38	0.76
对信息工程学院的提问	50	36	0.72
对宿舍情况的提问	50	40	0.80

根据表内准确率的描述，可以看出该项目的准确率基本符合，大约在 70%与 80%之间，能够基本满足用户的要求。

结论

基于 VSM 的聊天机器人微信小程序实现过程中，采用数据抓取工具进行学校问答数据的抓取，节省了部分数据收集的时间，但大部分数据采用人工收集的方式进行。在核心算法的实现过程中，严格按照软件工程进行需求分析、设计、实现、测试。在规定时间内用 VSM 算法、分词工具实现后台的搭建，用 request 实现微信小程序前台问答数据的发送，全过程实现了聊天功能。

该项目的功能主要是用户通过在微信小程序上进行提问，经过 VSM 算法将数据库对应数据进行识别并实现临时存储，将存储的数据计算 TF 值、IDF 值以及 Sim 值等数据，查找到最大值对应的问题，利用问题对应的 ID 进行数据库查找，从对应数据库中查到答案，返回给用户。整个过程能够让用户通过聊天快速获取阳光学院的信息，节省了用百度去查找数据的时间，为新生、家长、业界人士带来便利。

在基于 VSM 的聊天机器人微信小程序的开发过程中，利用微信小程序的开发与设计，实现了一个能够与用户交互的应用，能够让用户在此应用上快速获取信息；利用分词工具、VSM 算法和 TF-IDF 算法的综合应用，实现了后台问答数据的处理过程，后台的设计充分应用到自动问答系统的原理。总体来说，项目实现了基本聊天功能，让用户在查找阳光学院数据信息时，能够体会到与搜索引擎不一样的感受。

在实现过程中遇到一些问题，比如微信小程序与后台算法无法实现对接问题，通过 request 方法得到解决；接收数据出现问号乱码问题，通过修改发送数据的 header 得到解决等等。虽然解决了以上问题，但目前只能用开发者的微信账号进行问题的发送，数据量不足、数据处理的速度需要进一步修改与调整。

参考文献

- [1] 周蕾. 基于互联网的自动问答系统关键技术研究[实现][D]. 西安:西北大学软件工程专业, 2018.
- [2] 李舟军, 李水华. 基于 Web 的问答系统综述[J]. 计算机科学, 2017, 44(6):1-7.
- [3] 冯升. 聊天机器人问答系统现状与发展[J]. 机器人技术与应用, 2016(4):34-36.
- [4] 戴怡琳, 刘功申. 智能机器人的技术综述[J]. 计算机科学与应用, 2018, 8(6):918-929.
- [5] 刘佳雯. 语句相似度匹配在自动问答系统中应用与实现[D]. 南京:南京邮电大学电子与通信工程专业, 2018.
- [6] 曾文, 徐红皎, 李颖, 等. 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.
- [7] 高翔, 李志浩. Java Web 开发与实践[M]. 北京:人民邮电出版社, 2014:149-169.
- [8] 刘光瑞. Tomcat 架构解析[M]. 北京:人民邮电出版社, 2017:1-16.
- [9] 高洪涛. 从零开始学微信小程序开发[M]. 北京:电子工业出版社, 2017:2-20.
- [10] 王珊, 萨师煊. 数据库系统概论[M]. 北京:高等教育出版社, 2014:5-10.
- [11] Liu K, Zhao J, He S, et al. Question Answering over Knowledge Bases[J]. IEEE Intelligent Systems, 2015, 30(5):26-35.
- [12] 蔡礼高. 基于空间向量模型的文本内容筛查方法研究[D]. 成都:电子科技大学控制科学与工程专业, 2018.
- [13] 毛晓刚. 基于向量空间模型的本地搜索引擎的设计与实现[D]. 吉林:吉林大学软件工程专业, 2016.
- [14] 孙润志. 基于语义理解的文本相似度计算研究与实现[D]. 沈阳:中国科学院研究生院(沈阳计算技术研究所)计算机科学与技术专业, 2015.
- [15] 金镇晟. 基于改进的 TF-IDF 算法的中文微博话题检测与研究[D]. 北京:北京理工大学计算机科学与技术专业, 2015.
- [16] 谭静. 基于向量空间模型的文本相似度算法研究[D]. 成都:西南石油大学计算机应用技术专业, 2015.