硕士专业学位论文

调查问卷自动化审核系统的实现及其优化

Implementation and Optimization of the Questionnaire Automated Audit System

作者：刘翰文

导师：赵永祥

北京交通大学

2020 年 5 月

学位论文版权使用授权书

本学位论文作者完全了解北京交通大学有关保留、使用学位论文的规定。特授权北京交通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，提供阅览服务，并采用影印、缩印或扫描等复制手段保存、汇编以供查阅和借阅。同意学校向国家有关部门或机构送交论文的复印件和磁盘。学校可以为存在馆际合作关系的兄弟高校用户提供文献传递服务和交换服务。

（保密的学位论文在解密后适用本授权说明）

学位论文作者签名：　　导师签名：

签字日期：　　年　　月　　日　　签字日期：　　年　　月　　日

学校代码：10004 密级：公开

北京交通大学

硕士专业学位论文

调查问卷自动化审核系统的实现及其优化

Implementation and Optimization of the Questionnaire Automated Audit System

作者姓名：刘翰文 学 号：18125030

导师姓名：赵永祥 职 称：副教授

工程硕士专业领域：电子与通信工程 学位级别：硕士

北京交通大学

2020 年 5 月

致谢

本文的研究工作是在我的导师赵永祥副教授的悉心指导下完成的。赵永祥老师在我的研究过程中给予了我充足的帮助与耐心的指导。我在实验室攻读研究生期间，深深受益于赵老师的关心、爱护和谆谆教诲。作为一名优秀的导师，赵老师有着精益求精的学术精神、严谨的逻辑思维以及幽默风趣的表达方式和生活态度。能师从赵老师，我感到十分的荣幸。在此，谨向赵老师表示我最诚挚的敬意和感谢！

同时，我也要感谢所有教导过我，关心过我的实验室老师们。衷心感谢郭宇春老师、李纯喜老师、陈一帅老师、郑宏云老师、张立军老师、孙强老师和张梅老师在我的研究生学习阶段对我的指导与关怀，你们为我的学业倾注了大量的心血，你们为人师表的风范令我敬仰，严谨治学的态度令我敬佩。

其次，在实验室学习生活和撰写论文期间，实验室的苏迪师姐、胡玮师兄、刘一健师兄、宋云鹏师兄、张虎信、刘子可等同学对我的研究提供了帮助，在此向他们表示我的感谢之意。

最后，特别感谢父母在我求学生涯中给与我无微不至的关怀和照顾，一如既往地支持我、鼓励我。正是他们积极的鼓励和默默的奉献，才使得我顺利完成学业，成为社会的有用之才。

摘要

科普调查是国家掌握国民科学素质的重要工具，也是国家制定提升科学素质相关政策的重要依据。入户问卷调查作为科普机构采集各种科普调查数据的主要方式，调查过程的规范与否会影响获得数据的可靠性，进而影响相关决策的科学性。

目前对调查过程规范与否的审核主要依靠对于调查问卷信息进行人工审核，人工审核需消耗大量人力、物力资源，并且单份问卷的审核时间较长、效率较低，无法满足对逐年增长的调查问卷的审核需求。

采用调查问卷自动化审核系统来协助相关人员进行问卷的审核，可以减轻审核人员的压力，并提高审核问卷的效率。然而关于调查问卷自动化审核的研究相对较少，已有的部分研究中对于问卷审核也存在适用范围有限、无法提供错误原因解释等诸多问题。

为此，本文针对调查问卷可能出现的问题进行详细的分析，设计并实现了一套调查问卷自动化审核系统，从语音、图像和 GPS 等方面对问卷合格性进行审查，并使用 2019 年河北地区的调查问卷数据进行了测试。经测试，审核 1 份问卷的平均时间约为 1 分钟，效率远高于人工审核的效率。具体来说，本文主要工作包括以下几个方面：

(1)使用语音技术提高调查问卷的审核精度。通过语音识别技术，将调查过程中的录音信息转成文字信息，并根据相关题目的文本进行文本相似度计算，根据设定的阈值确定错误题目编号及类型，并使用静默检测、汉语拼音修正等方法提高系统的效率及准确率。使用音频分析对于问卷中各小题进行审核的 AUC 值达到了 0.95， Precision、Recall、F1 值分别达到了 0.96、0.93、0.94。

(2)使用图像识别技术和 GPS 技术提高调查问卷的审核精度。针对问卷中随机拍摄的图像，图像模块使用人脸检测方法识别图像中的人数和性别，以判断调查环境、受访者性别是否合乎要求。地理位置模块使用 GPS 功能对调查地点的地理位置进行分析，判断调查员是否按照要求前往了指定的居委会调查。数据表明图像模块可以审核出现有人工审核尚未发现的约 0.8%的性别作弊问题，地理位置模块对于现有人工审核出的地理位置作弊问题的召回率为 100%，且能够发现更多潜在的作弊问卷。

(3)在实际系统中实现了一套调查问卷自动化审核系统。对原有的后台管理系统进行了修改，实现了一套包含上述功能的调查问卷自动化审核系统并在阿里云服务器上进行了部署，提供了相关审核页面方便相关人员对错误原因进行查询。使用多线程并行处理、IP 地址检测等方法，保证了系统的有效性和可靠性。

关键词：文本相似度；问卷审核；系统设计；汉语拼音；静默检测

ABSTRACT

Science popularization survey is an important tool for government to master the scientific quality of the people, and it is also an important basis for the government to formulate related policies to improve scientific quality. The household questionnaire survey is the main way for science popularization agencies to collect various science popularization survey data. The standardization of the survey process will affect the reliability of the data obtained, and then the scientificity of related decisions.

At present, the review of the standardization of the survey process mainly depends on the manual review of the questionnaire information. Manual review consumes a lot of human and material resources, and the review time of a single questionnaire is relatively long and inefficient, which cannot meet the increasing survey Review requirements of the questionnaire.

The use of an automated questionnaire review system to assist relevant personnel in the review of the questionnaire can reduce the pressure on the reviewers and increase the efficiency of the review questionnaire. However, there are relatively few studies on the questionnaire automated review. In some existing studies, the questionnaire review also has many problems such as limited scope of application and failure to provide explanations for the causes of errors.

To this end, this article conducts a detailed analysis of the possible problems in the questionnaire, designs and implements a set of automatic questionnaire audit system, reviews the eligibility of the questionnaire from voice, image and GPS, and uses The questionnaire data was tested. After testing, the average time to review a questionnaire is about 1 minute, the efficiency is much higher than the efficiency of manual review. Specifically, the main work of this article includes the following aspects:

(1) Use voice technology to improve the accuracy of the questionnaire review. Through voice recognition technology, convert the recorded information in the investigation process into text information, and calculate the text similarity according to the text of the relevant topic, determine the wrong topic number and type according to the set threshold, and use silent detection, Chinese pinyin correction The method improves the efficiency and accuracy of the system. Using audio analysis, the AUC value of each sub-question in the questionnaire reached 0.95, and the Precision, Recall, and F1 values reached 0.96, 0.93, and 0.94, respectively.

(2) Use image recognition technology and GPS technology to improve the accuracy

of the questionnaire review. For the randomly captured images in the questionnaire, the image module uses the face detection method to identify the number and gender of the images to determine whether the survey environment and the gender of the interviewee meet the requirements. The geographic location module uses the GPS function to analyze the geographic location of the survey location and determine whether the investigator went to the designated neighborhood committee to investigate as required. The data shows that the image module can review about 0.8% of the gender cheating problems that have not been discovered by manual review. The geographic location module has a recall rate of 100% of the geographic location cheating problems that have been manually reviewed and can find more potential cheating Questionnaire.

(3) In the actual system, a set of automatic questionnaire audit system is implemented. The original background management system has been modified, and a set of automatic questionnaire audit system containing the above functions has been implemented and deployed on the Alibaba Cloud server. Related audit pages are provided to facilitate relevant personnel to query the cause of the error. The use of multi-threaded parallel processing, IP address detection and other methods ensure the effectiveness and reliability of the system.

KEYWORDS：Text similarity；Questionnaire review；System design；Hanyu Pinyin；Silent detection

目录

1 引言

1.1 研究背景及意义

科普调查又称公民科学素质测评，是国家掌握国民科学素质的重要工具，也是国家制定提升科学素质相关政策的重要依据。为了给科学知识在社会中的传播提供基础，让公众及时有效地跟踪科学发展、了解科学报道和提升公民的科学素质，科普相关机构需要采集各种各样的科普数据资料进行分析与研究[1]。

我国某著名科普理论研究机构（以下简称“科普机构”）近几年来每年都开展了“公民科学素质测评”工作，通过收集科普数据资料并进行分析与研究，形成了一套稳定而成熟的中国公民科学素质测评体系，承担着我国公民科学素质检测评估与科学决策的历史重任。

调查问卷作为该科普机构采集各种各样科普数据资料的重要调查工具，其调查结果在整个中国公民科学素质测评体系中起到了举足轻重的作用。

因为调查结果是决策的重要依据，该科普机构指定了严格的流程保证调查结果的准确性。其进行科普问卷调查时的流程如下，首先，根据科普调查的需求对问卷进行设计[2]并聘用了专业的团队设计了用于安卓平板上的调查问卷 app 并建立了对应的后台系统。随后，科普机构与专业调查机构进行合作，培训出了一批可以从事问卷调查的调查员。这些调查员携带者装有调查问卷 app 的平板并根据科普机构的安排，前往对应的居民小区进行抽样调查。调查全过程以录音和图片的形式保存，结果以编码方式被保存在系统后台以便后续整理与分析。最后，将所有合格问卷的数据进行统计，移交给相关科普研究小组进行数据分析。

尽管科普机构制定了严格的操作规范来确保最终获得高可靠度的数据，但仍然不能避免不合格问卷的出现。一方面是由于所得绩效与完成调查问卷数量相关，调查公司的调查员因为利益的驱使，不按照相应的“科普调查规范”进行调查，出现错读、漏读、不读、催促或误导受访者进行答题的现象。另一方面是由于部分地区的调查难度大，且对受访者年龄、性别有严格要求，出现调查员不按要求随便找人应付调查的情况。此外，也有受访者敷衍了事、不配合的情况产生。

为此，该机构采用人工审核的方法解决问卷质量问题。上述原因所导致的最终获得的数据可靠性大幅度降低，据统计错误样本的数量占到了总样本数的 20%-30%。为此，科普机构联系专业的第三方审核公司进行两轮问卷人工审核，审核公司根据“科普调查规范”对所有问卷进行人工审核，将不合格的问卷从数据库中标

出，不纳入最终的统计与分析。

但是实践证明，人工审核仍然面临严重的挑战。由于人工审核受主观成分影响较大，且单份问卷的录音时间较长，审核人员无法自觉做到全部按要求完成审核。再加上问卷的数量过于庞大（通常是数十万数量级的），且调查及审核时间只有 3-4 个月，想要按时高质量完成需要投入大量的人力物力，在现在高额的人工费用的情况下，成本巨大。因此亟需一种自动审核的方法协助审核人员进行审核，增加审核效率，起到缩短审核时间、降低用人成本的作用。

在自动审核方面，目前比较的缺乏这方面的研究。近年来，随着人工智能的兴起以及算法浪潮，语音识别[3]、人脸检测[4]、图像识别[5]以及传统的自然语言处理等技术越来越成熟，使问卷自动审核成为了可能。

综上，建立一套可用于审核问卷的自动化审核系统，协助审核人员高质高效地完成审核工作，为做大做强科普调查与宣传这种社会公共服务提供帮助，具有深远的社会意义，同时也是对人工智能及自动化等现代科技在科普调查这种传统工作上应用的一种探索，证明其应用价值。

本文的研究工作是设计一款能够协助科普机构审核调查问卷是否合格的自动化审核系统，通过语音识别将问卷中的语音部分转成文字，再用文本相似度分析、人脸检测等方法结合科普机构的“科普调查规范”确定出不合格问卷，并指出具体的错误位置供科普人员参考。

1.2 国内外研究现状

本文的研究工作是设计一款能够利用语音、图像和文字等技术实现调查问卷合格性的的自动化审核系统，涉及到问卷可信度检测、语音识别等领域，下面介绍这些领域的研究现状。

1.2.1 调查问卷可信度研究现状

调查问卷在科普机构的调查进行过程中起到了至关重要的作用，因此，为了减少和避免潜在的差错，提高科普实践活动的效果，提升科普知识理论与应用研究的水平，必须在调查过程中获得具有高可靠性的数据。

目前的研究中，一种方法是从问卷的主要内容出发，根据探究和掌握调查问卷及其设计要素的方式，提高调查问卷获取可靠数据的水平。调查问卷的问题分为四个类型，背景性的问题、客观性问题、主观性问题和检验性问题。其中检测性问题就是为了检测填答是否真实、准确而设计的。问卷通过将此类问题分散安插在问卷

的不同位置，并通过这些问题之间的逻辑设计来相互检验，判断受访者的回答是否自身矛盾。然而，对于检测性问题的设计十分繁杂，且检验出填答存在疑义的效果也有待商榷额，只依靠设置检验性问题的方法无法很好的解决提高问卷可靠性的问题。

另一种方法是交由审核公司进行人工审核，审核人员经过相关的培训，通过科普机构指定的“科普调查规范”，根据录音、照片等信息对问卷逐一进行检测。同时，审核人员还会通过以往相关工作的审核经验去判断该份问卷是否合格，如根据后台统计的调查人员的以往可信度、音频中出现的特殊事件、照片信息是否与统计信息一致、GPS 统计信息与调查地的具体情况等。

随着现代社会智能化、信息化的发展，机器学习、人工智能等科技越来越多的应用，诞生了使用机器学习进行调查问卷审核的方法。苏迪在文献[6]中提出了根据问卷所提供的信息提取出高维度特征，并使用不同的机器学习模型，如逻辑回归、决策树、XGBoost 等方法结合特征对调查问卷的可信度进行打分，设置阈值判断问卷是否合格。苏迪的方法在一定程度上取得了不错的效果，但无法定位具体错误原因或是错误位置，使得科普机构对该方法的信任程度不高。

1.2.2 语音识别现状

本文将采用语音识别技术分析调查问卷中的语音，下面简要介绍这方面的现状。

语音作为最简洁自然的交流方式，一直是人机交互中的重要研究领域之一。语音识别作为实现人机交互的关键技术，目的是使计算机可以“听懂”人类的语言，能将语音转换成文本。近年来，随着信号处理、人工智能等领域的发展，使得语音识别在处理效率和处理精度都取得了显著的成效。日常生活中我们所使用的苹果的 Siri，微软的小娜和科大讯飞的语音输入法都是语言识别技术方面的佼佼者。

传统的语音识别技术是基于 GMM（高斯混合模型）-HMM[7]（隐马尔科夫模型）框架的。其核心是用隐马尔可夫模型 HMM 对语音时序进行建模，再用高斯混合模型 GMM 对语音的观察概率进行建模。上个世纪 90 年代，随着语音识别声学模型的区分性训练准则和模型自适应方法的提出，语音识别取得了不错的发展。

2006 年 Hinton 提出 DBN[8]（深度置信网络），促使了 DNN[9][10]（深度神经网络）研究的复苏，掀起了深度学习的热潮。人们开始探索利用神经网络替代 GMM 进行语音识别。语音识别需要对波形进行加窗、分帧、提取特征等预处理，训练 GMM 时候，输入特征一般只能是单帧的信号，而对于 DNN 可以采用拼接帧作为输入，这些是 DNN 相比 GMM 可以获得很大性能提升的关键因素。但是由于 DNN

输入的窗长是固定的，学习到的是固定输入到输入的映射关系，从而导致 DNN 对于时序信息的长时相关性的建模是较弱的。考虑到语音信号的长时相关性，RNN[11][12]（循环神经网络）这种拥有长时相关性与记忆功能的网络更适合进行语音识别。而 LSTM（长短时记忆模块）的引入解决了传统 RNN 梯度消失的问题，使得 RNN 框架在语音识别效果上超越了 DNN，成为了当今语音识别的主流框架。此外，研究人员还加入了深层双向 RNN 使得使用历史语音的同时还可以使用未来的语音信息进行决策，提高决策的精度。

本文使用了科大讯飞的语音转写功能[13]，该功能用到了科大讯飞自研的 DFCNN（深度全序列卷积神经网络），DFCNN 结合了科大讯飞自研的 FSMN[14]（前馈型序列记忆网络）与 RNN（卷积神经网络）的优点，使得将语音识别延迟降到了 180ms 内的同时比双向 RNN 框架的性能提升了 15%。

1.3 本文研究内容

本文旨在设计一套调查问卷的自动化审核系统，使用了语音识别、文本相似度检测、人脸检测等方法，来协助科普工作人员对收集到的调查问卷进行审核，具体工作包括以下几个方面：

（1）通过语音识别技术，将调查过程中的录音信息转成文字信息，并根据相关题目的文本进行文本相似度分析，根据设定的阈值确定错误题目编号及类型，并使用静默检测、汉语拼音修正等方法提高系统的效率及准确率。

（2）针对问卷中随机拍摄的图像，使用人脸检测方法得出图像中的人数，并判断调查环境是否合乎要求。对检测出的人脸进行性别识别，判断调查员是否找到了对应性别的受访者进行调查。使用 GPS 功能对调查地点的地理位置进行分析，判断调查员是否按照要求前往了指定的居委会调查。

（3）为了保证系统可以实际运行，实现了包括上述功能的问卷审核系统，对原有的后台管理系统进行了修改，并在阿里云服务器上部署该问卷审核系统，并提供了相关审核页面方便相关人员对错误原因进行查询。使用多线程并行处理、IP 地址检测等方法，保证了系统的有效性和可靠性。

1.4 论文结构安排

本文总共分为六个章节，每个章节阐述的内容如下：

第一章为引言部分，重点介绍了调查问卷自动化审核系统的研究背景与意义，以及相关研究的研究现状。最后概述了本文的研究内容与结构安排。

第二章为相关研究及技术介绍，主要介绍了中文分词、字符串相似度匹配、以及人脸检测算法等方面的相关技术。

第三章为需求分析和系统总体设计，首先对调查问卷自动化审核方面的需求进行了分析，随后介绍了系统以及系统内各个模块的总体设计，最后提出了有关系统的优化方法。

第四章为模块的具体设计部分。主要对接口的设计与部署，原有系统中定时脚本的设计，审核算法的音频、图像、地理位置模块进行了介绍，详细说明了具体的实现方法。

第五章为系统测试结果及分析部分，先对实验所使用的数据进行了统计与分析，然后展示出了本文使用检测算法的效果，最后对系统的性能进行了测试。

第六章为总结与展望部分，主要总结了本文的工作内容，并对后续工作的工作方向提出了展望。

2 相关研究与技术

本章对后续章节中调查问卷自动化审核系统的研究所用到的相关技术进行了介绍。共分为 4 个小节，第一节介绍了文本分词的相关研究，第二节介绍了文本字符串相似度匹配的相关研究，第三节介绍了人脸检测算法的相关研究，第四节介绍了性别检测的相关研究，第五节为本章小结。

2.1 文本分词相关研究

早期对文本进行处理及分析的技术多产生与国外，涉及到的语言多为英语或其他各种语言，这类语言的特点之一就是词与词之间有空格作为明确的边界，因此研究人员可以根据空格快捷高效地为一段文字中的每句话进行分词处理，从而进行后续的操作。不同于这些语言，中文文本中的词与词之间没像空格这样的明确边界作为划分，汉字不同于英语中的单词，我们不能简单的将一句话中的每个汉字提取出来单独分析，“词”才是句子的基本组成单位，要对文本中的句子进行分析，我们必须确定每个词的边界，将一句话切分成若干个一字或多字词语，作为后续其他自然语言处理操作的基础。目前现有的分词技术大致可分为三种：基于字符串匹配的分词方法[15]、基于统计的分词方法[16]、基于词义的分词方法[17]。

(1) 基于字符串匹配

最简单的基于字符串匹配方法是 MM（正向最大匹配算法），算法思想为：按照文本的阅读顺序，从左至右截取一定长度的汉字（长度为最长的词语字数），与已有的词典中的词语进行一一比对，如所有均无法匹配则缩短一个长度再进行匹配，直到匹配成功为止。如果匹配成功就把该汉字作为一个词切出来，并对剩余的文本重复这样的操作直至所有词被切分出来。由于这种分词方式以字典中的词作为基础，因此当出现一些歧义或是字典里没有的新词或是人名、地名时，效果并不理想。

（2）基于统计的分词方法

典型的如组合度算法，该方法利用词频统计的方法根据上下文中相邻字出现的概率来判断，当概率高于一定值得时候就将这些相邻字作为一个词看待，这个统计方法对切分歧义词和识别新词有着良好的效果。但这种方法时常会切分出一些无用的词语，对于那些词频较低的词语也无法切分出来。

（3）基于词义的方法

典型的如神经网络法[18]，使用神经网络将如 LSTM 等在分词。采用有监督学

习的方式，通过对大量人工分词文本进行训练，使神经网络学习到文本中句法、语义信息，利用这些句法语义信息来进行词义标注，已解决分词中歧义的现象。由于汉语的语法规则十分笼统、复杂，且进行人工分词标注时造成了大量的人力消耗，因此这种分词的效果无论是在时间成本、时间复杂度方面还是在分词准确度方面都很难达到令人满意的效果。

目前在分词领域有很多现有的研究成果，多数为开源的程序，如 HanLP、结巴分词[19]、哈工大的语言技术平台[20]等。本文中编程使用的开发语言为 Python，因此综合考虑使用了结巴分词这一比较成熟的开源分词工具。

2.2 字符串相似度相关研究

在自然语言处理的过程中，我们通常会遇到需要找出相似语句的场景，亦或是找出句子的相近表达，这里面就涉及到了字符串的相似度计算的问题。字符串相似度算法是指通过一定算法，来计算两个不同字符串的相似程度。通常会用一个百分比来衡量字符串之间的相似程度[21][22]。下面介绍几种字符串相似度分析相关的技术。

2.2.1 编辑距离算法

最常见的字符串相似度算法是编辑距离算法（EditDistance），又称 Levenshtein 算法[23]。该算法将两个字符串的相似度问题，归结为其中一个字符串转化成另一个字符串所需要进行的最少的编辑操作次数。编辑操作次数越高，就说明两个字符串的相似度就越低。许可的编辑操作包括，将一个字符替换成另一个字符，插入一个字符，删除一个字符。

例如我们拥有两个字符串 strings 与 setting，我们想讲 string 转化成 setting，那么我们需要进行如下操作：

第一步，在 s 和 t 中间加入 e。

第二步，将 r 替换成 t。

第三步，删除结尾的 s。

所以 strings 与 setting 的编辑距离为 3，这就意味着两者相互转化所要改变（添加、替换、删除）的最小步数为 3。

2.2.2 Jaccard 相似度

Jaccard 相似度系数，用于比较有限样本集之间的相似性与差异性。Jaccard 是一种常用的计算样本相似度的方法，主要可以应用场景为字符串相似度计算、数据聚类、文本去重与查重、计算对象间的距离等。Jaccard 的相似度系数值越高，说明两个样本就越相似。

公式如下：

$$J(A,B)=\frac{|A\cap B|}{|A\cup B|}=\frac{|A\cap B|}{|A|+|B|-|A\cap B|} \quad (2-1)$$

$$d_j=1-J(A,B)=1-\frac{|A\cap B|}{|A|+|B|-|A\cap B|} \quad (2-2)$$

其中 J(A,B)∈(0,1)。

例如“今天心情很好”与“今天心情不错”这两个词，他们的交集为（今、天、心、情），并集为（今、天、心、情、很、好、不、错），因此他们的 Jaccard 相似度为 0.5，Jaccard 距离为 1-0.5=0.5。

2.2.3 Word2Vect / TF 结合余弦距离算法

Word2Vect 可以用于计算句子之间的文本相似度[24]，它是 Google 于 2013 年开源推出的一个用于获取 word vector 的工具包。Word2Vect 算法的基本思想是通过语料库的大量文本作为输入，通过训练将每个不同的单词转化为固定长度的词向量。不同于传统的布尔编码（One-hot representation）将一个词用一个很长的向量的 0 或 1 值表示（向量长度为所有构成样本的词去重后的词典大小），Word2Vect 属于分布式表示方法，它可以将单词转化为相对较小的向量，并且向量中的数值不再是 0 或 1，而变成了浮点数。Word2Vect 有 Skip-gram (Continuous Skip-gram Model) 和 CBOW[25] (Continuous Bag-of-Words)两种训练方式，下面介绍这两种训练算法的模型：

由图 2-1 可知，Skip-gram 的模型分为三个部分，输入层、隐式映射层、输出层。

输入层：首先我们先通过布尔编码即 One-hot 编码将一个词转化成 One-hot 向量。假如我们的词典中有 10000 个互不相同的词语，那么我们的向量维度就是 10000 维，对于每个词语而言，词语在词典中出现在第 N 个位置，那么该词语所对应向量在第 N 个维度取 1，其余位置取 0。这样我们就拥有了所有词语对应的向量。由上图可知，我们需要从样本中取出一个词将其转化为向量后作为输入数据 $w(t)$。

隐式映射层：如果我们希望最终使用 300 维度的特征来表示一个单词，那么的隐式映射层应有 300 个节点，权重应该为 10000 行，300 列。这样我们就可以通过

这个矩阵，将 10000 维的向量转成我们需要的 300 维向量，我们最终的目标就是通过训练让神经网络学习这个隐式映射层的矩阵。

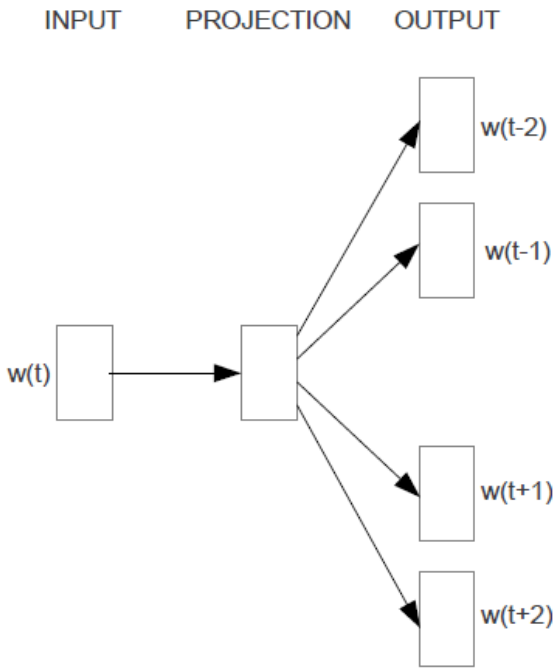


图 2-1 Skip-gram 模型结构图[25]

Figure 2-1 The structure of Skip-gram model[25]

输出层：由上图可知，当我们从样本中取一个词语作为输入时，需要从这个词语的前后各取出两个词 *w（t-2）*、*w（t-1）*、*w（t+1）*、*w（t+2）*作为输出层的目标。输出层实际上是若干个（这里是 4 个）softmax 分类器，每个分类器的每个节点会输出一个 0-1 之间的概率值，所有输出节点的概率值之和为 1。我们需要通过训练使分类器的输出结果与希望输出层输出的目标尽可能一致（目标对应的概率接近 1，其余位置为 0），这样我们就可以在训练中不断的完善并得到我们想要的隐式映射层参数。

CBOW 模型与 Skip-gram 模型的原理相同，只是网络结构相反。

由图 2-2 可知，CBOW 模型也具有输入层、隐式映射层、输出层三层构成，其输入输出层的 One-hot 编码与 Skip-gram 模型相同，只是输入的数据与输出的数据相互对调，在输出层之前需要将 *w（t-2）*、*w（t-1）*、*w（t+1）*、*w（t+2）*转成的四个对应的 300 维的特征编码相加再经过输出层得到对应目标。

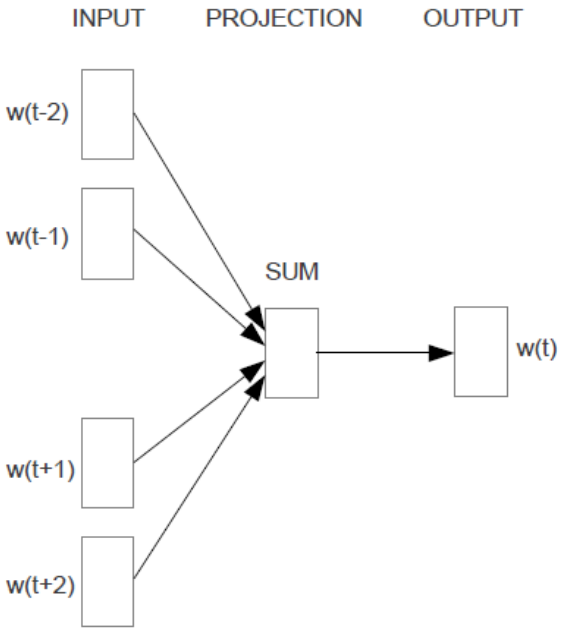


图 2-2 CBOW 模型结构图[25]

Figure 2-2 The structure of CBOW model[25]

经过上述两种 Word2Vect 的方法，我们将文本中的字符串中的词语进行切分，并将每个词语转变成对应的词向量。我们将每个字符串对应的词向量相加，得到对应的两个词向量。计算两个向量的相似度一般用到的方法是余弦相似度算法，该算法的公式如下：

$$cos\theta = \frac{A \cdot B}{||A||B||} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n}(A_i)^2} \times \sqrt{\sum_{i=1}^{n}(B_i)^2}} \qquad (2-3)$$

余弦相似度范围从-1 到 1，-1 意味着两个向量截然相反，1 表示他们完全相同。这里余弦相似度的值越接近 1，说明两个字符串就越相似。

同理，我们也可以使用 TF 算法，将字符串切分字或词语并转化成 TF 矩阵，如“你在干嘛呢”和“你在干什么呢”，先获取词表内容['么', '什', '你', '呢', '嘛', '在', '干']，再将两个字符串转成对应的向量矩阵[0 0 1 1 1 1 1]与[1 1 1 1 0 1 1]。再利用余弦相似度计算两个字符串之间的相似程度，这里由于词频统计的值为非负整数，因此利用词频计算出的余弦相似度的取值在 0-1 之间，值越大说明字符串相似度越高。

2.3 人脸检测算法相关研究

人脸检测算法是指按照一定的策略将人脸及五官的位置从图像或者视频背景中检测出来。由于其在在视频会议、安全访问控制、美颜相机等众多领域具有非常重要的应用价值，近年来成为了人工智能领域与计算机视觉领域炙手可热的研究话题。

早期的人脸检测算法使用的是模板匹配技术，即预先获得一个人脸模板图像，再到被检测图像中的各个位置进行匹配，确定这个位置是否有人脸。机器学习算法兴起后，支持向量机、神经网络等算法被运用到了这种传统的人脸检测算法中，该算法可以当成对图像上某个区域进行的人脸-非人脸的二分类判别。最初该算法只能用于解决正面的人脸检测问题，后来通过在二分类判别之前加入一个神经网络估计人脸的旋转角度从而解决了多角度的人脸识别问题。但受限于算法的思想，该算法并不能很好的解决人脸检测问题。

第二代人脸检测算法基于 AdaBoost 算法与反映图像灰度变化的 Haar-like 特征。在 2001 年 Viola 和 Jones 设计了一种人脸检测算法 VJ 算法。它使用简单的 Haar-like 特征和级联的 AdaBoost[26] [27]分类器构造检测器，检测速度较之前的方法有 2 个数量级的提高，并且保持了很好的精度。VJ 算法较好的解决了近似正面人脸的检测问题，此后研究人员根据算法原理提出大量改进方案，包括拓展的 Haar 特征以及 ACF 特征等，在深度学习算法出现之前，一直是人脸检测算法的主流框架。

随着深度学习框架的发展，卷积神经网络在图像分类问题上取得的成功很快被研究人员运用到了图像检测问题上。其精度远远超越了之前的 AdaBoost 算法。鉴于直接用滑动窗口进行图像分类的人脸检测算法计算量巨大难以实现实时计算，使用卷积神经网络进行人脸检测的方法采用了很多手段来解决计算过大这一问题。

如今已经有一些高精度、高效的算法被研究出来，如 Cascade CNN、Face R-CNN[28]、MTCNN[29]等。

由于本文主要用到了 MTCNN 的方法，接下来主要介绍一下 MTCNN 的算法结构。

MTCNN 算法包含三个子网络，Proposal Network(P-Net)、Refine Network(R-Net)、Output Network(O-Net)，这三个网络对人脸的处理依次从粗到细，具体的算法结构如图 2-3 所示：

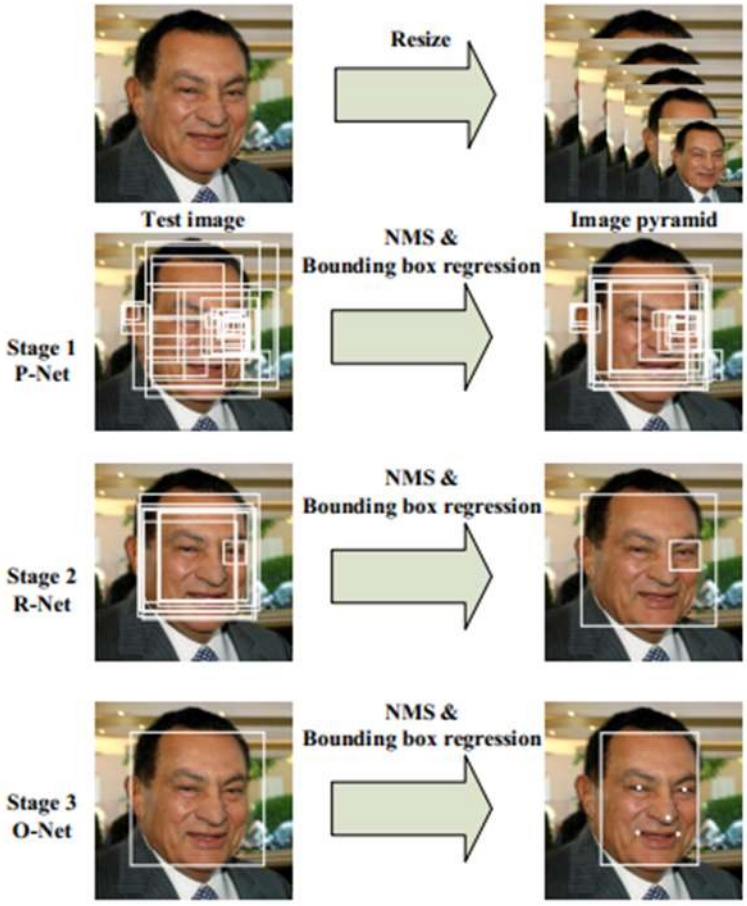


图 2-3 MTCNN 算法结构图[30]

Figure 2-3 MTCNN algorithm structure diagram[30]

使用这三个网络之前，需要先对图像进行预处理。使用图像金字塔将原始图像缩放到不同的尺度，然后再将不同尺度的图像放入到这三个网络中进行训练，这样做的目的是为了可以检测到不同大小的人脸，从而实现多尺度的检测。

（1）P-Net

P-Net 的主要目的是为了生成一些候选框，我们通过使用 P-Net 网络，对图像金字塔图像上不同尺度下的图像的每一个 12×12 区域都做一个人脸检测。

P-Net 的输入是一个 12×12×3 的 RGB 图像，在训练的时候，主要使得网络判断这个 12×12 的图像中是否存在人脸，给出人脸框的回归和人脸关键点定位并对每个可能存在人脸的图像进行打分。

在进行人脸检测的时候输出只有若干个个边界框的 4 个坐标信息对应的 Score，当然这 4 个坐标信息已经使用网络的人脸框回归进行校正过了，对应的 Score 可以看作是分类的输出(即人脸的概率)。

由于 P-Net 的输入是图像金字塔，每张图片都需要 reshape 导致处理时间较长，因此 P-Net 在检测时进行的比较粗略。

（2）R-Net

R-Net 与 P-Net 类似，为了解决 P-Net 在粗略检测时可能产生的问题，要使用 R-Net 进行进一步的优化。这一步的输入是上一步 P-Net 输出的边框，在进入 R-Net 前，无论实际边框有多大，都要缩放到 24×24×3 的大小。

R-Net 的最后一层是两个全连接分支，一个输出人脸概率的预测向量，是对 P-Net 预测的在再次筛选，另一个是输出 4 维的位置向量，目的是对输入的人脸坐标框进行再次调整。

（3）O-Net

进一步将 R-Net 所输出的候选区域缩放到 24×24×3，然后输入到 O-Net 进行最后一步计算。O-Net 的基本结构是一个较为复杂的卷积神经网络，相对于 R-Net 多了一个卷积层，这也意味着检测效果会更加优秀。O-Net 的最后一层有三个全连接层，前两个是输出对人脸框的打分及调整，第三个输出检测到的人脸五官的五个坐标。

P-Net、R-Net 与 O-Net 都是使用了 NMS（极大值抑制）的方法对高度重叠的候选窗口进行了合并，以确保同一张人脸不会被框出来两次。

2.4 性别检测相关研究

自 2012 年以来，深度学习获得了越来越多的关注，更多的精英学者投身到了对于深度学习的各方面研究中，作为深度学习在现实中应用的一个重要分支，性别检测也获得了迅速的发展。AlexNet，vgg16，vgg19，gooGleNet 等各种模型层出不穷，使得性别检测在速度和精度方面都获得突破性的进步。

在解决图像识别问题上主要使用的是卷积神经网络 CNN[5]，它可以使用卷积核，在训练过程中对图片进行自动特征提取。卷积神经网络直接用原始图像的全部像素作为输入，但是内部为非全连接结构。因为图像数据在空间上是有组织结构的，每一个像素在空间上和周围的像素是有关系的,和相距很远的像素基本上是没什么联系的，每个神经元只需要接受局部的像素作为输入，再将局部信息汇总就能得到全局信息。

这种做法中比较经典的是，先将输入的图片数据通过多个不同的卷积层提取人脸特征，再通过几个全连接层处理这些特征，其中也使用了 drop 层和 norm 层提高泛化能力和防止过拟合，最终通过 sigmoid 层或是使用 SVM 等传统机器学习方法对数据进行最终的性别分类。

权值共享和池化两个操作使网络模型的参数大幅的减少，提高了模型的训练效率。权值共享指的是，在卷积层中可以有多个卷积核，每个卷积核与原始图像进行卷积运算后会映射出一个新的 2D 图像，新图像的每个像素都来自同一个卷积

核。池化其实就是降采样，对卷积(滤波)后，经过激活函数处理后的图像，保留像素块中灰度值最高的像素点(保留最主要的特征)，比如进行 2X2 的最大池化，把一个 2x2 的像素块降为 1x1 的像素块，这样做可以降低模型的计算量，从而高效的完成任务。

2.5 本章小结

本章对调查问卷自动化审核系统所用到的相关技术进行了介绍。首先介绍了中文分词的三种模式，然后介绍了使用编辑距离、Jaccard 相似度、余弦距离和 Word2Vect 进行文本的字符串相似度匹配，接着介绍了人脸检测技术，重点介绍了目前流行的基于卷积神经网络的 MTCNN 人脸检测技术。最后介绍了使用卷积神经网络进行的性别检测。本章所介绍的技术，是实现本文算法设计的技术基础。

3 需求分析和系统设计

本章给出了的调查问卷自动化审核系统的整体设计和系统中各模块整体结构。共分为 8 个小节，第一节给出了需求分析，第二节给出了系统的整体设计，第三节设计了原有的系统与审核算法的接口流程，第四到六节分别设计了语音分析、图像以及地理位置模块，第七节给出了两个系统性能优化的方法，第八节为本章小结。

3.1 需求分析

3.1.1 背景介绍

本节主要介绍了调查问卷的调查流程、内容组成、问题的结构以及问卷审核现有的研究进展。

调查问卷是科普机构进行科普调查的重要手段，其主要的调查流程为：(1)科普机构设计专业的问卷；(2)聘请调查员携带平板电脑针对指定地区用户进行入户调查；(3)通过后台管理系统对问卷的可靠度进行评估；(4)统计合格问卷的数据信息，进行科学的分析与研究。

调查问卷主要由导语、科普调查题目和结束语组成，其中科普调查题目包含四个部分，分别为：被访者的基本信息、公民的科技信息来源、公民的科学的理解和公民对科学技术的态度。

调查问卷的每道问题由题干、编号、题项和选项组成，以“公民的科技信息来源”部分的 B2 题为例，具体信息如图 3-1 所示：

B2. 您对下列科技发展信息是否感兴趣？感兴趣的程度如何？（每行选一项）

题 项	非常感兴趣	一般感兴趣	不感兴趣	不知道
(1) 宇宙与空间探索	○	○	○	○
(2) 环境污染及治理	○	○	○	○
(3) 计算机与网络技术	○	○	○	○
(4) 遗传学与转基因技术	○	○	○	○
(5) 纳米技术与新材料	○	○	○	○
(6) 新能源开发及利用	○	○	○	○

图 3-1 公民科学素质抽样调查问卷 B2 题

Figure 3-1 Sampling questionnaire for citizen scientific quality Question-B2

在实际调查中，调查员为了节约成本（时间、交通费等）以谋取更多利益，可能会在调查过程中作假。例如：(1)为了缩短调查时间，调查员读题目时故意漏读部

分内容，有时甚至不读题目，让受访者直接勾画答案。(2)对于一些偏僻或是调查难度大的“地区 A”，有些调查员选择前往附近相对容易调查的“地区 B”进行调查，然后将得到的“地区 B”的数据打上“地区 A”的标签进行提交。(3)调查员为了使受访者的男女性别数量与任务上一致，故意将某些受访者的性别信息从“男”改成“女”或从“女”改成“男”。

为了保证调查的可信度，需要对问卷进行审核。除了传统的人工审核外，苏迪在文献[6]中提出了一种基于机器学习的调查问卷可信度研究方法。该方法从音频、图像等方面对调查问卷提供的数据进行了数据挖掘并筛选了特征，使用了几种经典的机器模型进行对比，找到了现有数据集上表现最优的 XGboost 作为最终的模型为问卷的可信度进行打分。

3.1.2 需求分析

本节中，首先对已有的审核算法的缺点和需要改进的问题进行了分析，后介绍了原有的调查问卷系统并指出了文本设计需要对该系统进行改进的地方。

本文在与客户（科普机构）长期沟通和调研的基础上，在本课题组苏迪[6]的研究基础上进行了需求分析，总结如下：

在音频方面，针对调查员在调查过程中为了节约时间，缺读、漏读题目以及语速过快等问题，导致问卷不合格。苏迪的工作使用了音频功率高低来判断某一时刻是否存在语音，进而计算出了调查员朗读单个问题的语音总时长以及静默时长的序列，并把对应的数值作为特征输入到机器学习模型中进行训练。但是，这种方法忽略了音频中最重要的语义部分，因此损失了大部分有效信息，无法对错误进行解释。例如，有些音频中，调查员一边和受访者聊天，一边让受访者自己填写问卷，苏迪的算法无法区分对话内容是否与问卷相关，而会因为的确有说话声音的存在而把这份本应不合格的问卷判为合格。

在图像方面，调查问卷要求调查进行在调查员与受访者一对一的环境中，比如入户调查。但调查员为了节约时间，可能会前往人员密集的公共场所（环境中存在多人）进行调查，或是调查员自己模拟受访者声音，自己提问自己回答（环境中只有一人），从而导致问卷不合格。为了解决这个问题，苏迪利用了调查问卷的 APP 在调查过程中随机拍摄的照片，对图中出现的人数进行了识别，如果出现人数不为两人，则会降低问卷的可信度。但仍面临如下问题：(1)只是对人数进行了识别，而没有对相应的受访者性别进行识别，失去了性别这一有效信息；(2)苏迪使用的 YOLO[31]人脸检测网络过于庞大，无法在缺少 GPU、算力受限的环境下使用。

在地理位置方面，由于一些指定需要调查的“居委会 A”位置偏僻或是调查难

度较大，部分调查员为了完成调查“居委会 A”的目标，前往调查相对轻松的“居委会 B”进行调查，后将所得数据打上“居委会 A”的标签。苏迪的方法根据 GPS 信息计算了目标居委会与调查地点的距离，并设置了统一的阈值对全国所有地区的居委会进行判定。但是，这种方法忽略了我国各地区之间地域差异较大的国情。有些地区居委会管辖面积大（如西藏、新疆等地），有些居委会管辖面积小（如北京城区），使用单一阈值进行判定可能会产生误判或漏判。

在实际的系统应用方面，本文对实际系统的实现进行了完善。苏迪的论文给出了算法，但是该机器算法是在 windows 下的本地环境上运行的，没有嵌入到实际的工作系统中。本文把机器识别模块嵌入到了实际的工作系统中。

最后，苏迪的论文为调查问卷提供了一个可信度的数值型结果，本文算法进一步给出了不合格问卷的错误类型和具体的错误位置，使检测结果更具有说服力。

总的来说，本文将实现如下需求：

（1）音频方面，本文先将调查问卷中的语音识别成文本，然后针对每道问题的具体内容进行文本相似度分析，设置阈值对相似度的值进行判定，判断问卷是否合格。这样做可以充分利用音频中的语义信息，提高了审核的准确性，并能够定位出错误的类型和具体错误位置。

（2）图像方面，本文使用了运算量较小的 MTCNN 人脸检测网络，并自己搭建了 CNN 网络进行性别二分类计算。这样做的好处是，可以将图像模块正常部署在算力较低的服务器上并以较快速度运行，而且能够通过性别识别发现调查员的作弊行为。

（3）地理位置方面，本文不仅计算了调查地点与目标居委会的距离，还计算了调查地点与附近所有居委会距离，通过这些距离对地理位置问题进行分析。相比于苏迪只用一个距离和单一阈值进行判断，这样做更为灵活、准确率更高且能够给出错误前往的居委会名称，便于后续的核实。

（4）实际系统应用方面，本文需要对原有系统进行改进并设计连接算法模块的接口。具体需求为：在原有系统中设置定时任务，周期性从数据库读取数据并调用接口连接算法模块得到审核结果，后在原有系统中添加前端页面进行结果展示。这样的系统可以自动对所有问卷进行机器审核，并能够使科普人员看到最终的审核结果。

由于已经有一个调查问卷系统，本文需要设计一个机器审核系统，并将这个机器审核系统加入已有的调查问卷系统中，本文设计的机器审核系统与调查问卷系统之间的关系如图 3-2 所示：

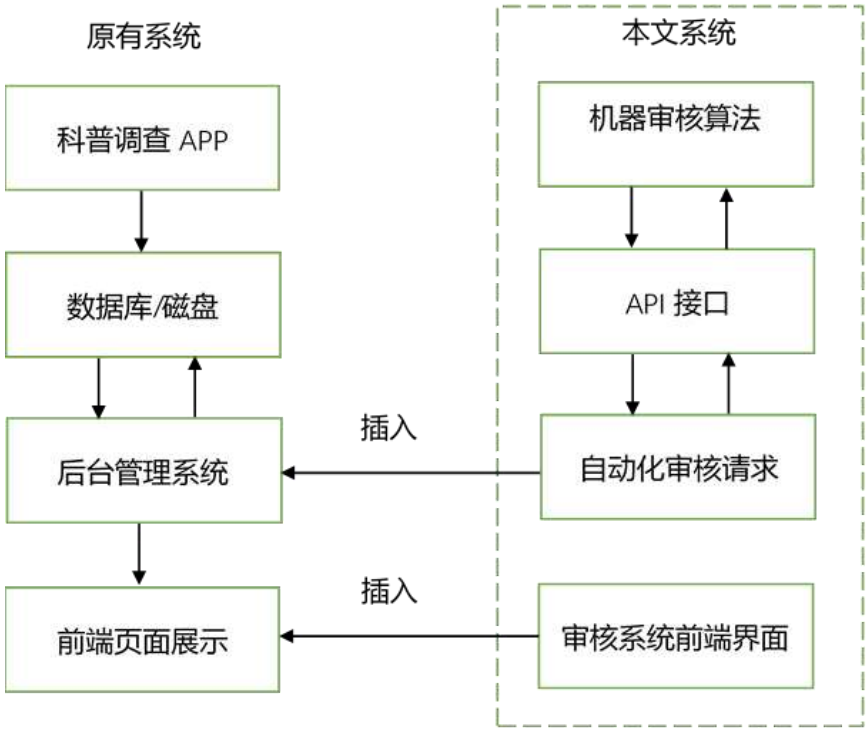


图 3-2 原有系统与本文系统介绍

Figure 3-2 Introduction of the original system and this system

从图中可以看出，原有的系统由 4 部分组成，首先调查员使用科普调查 app 进行科普调查，然后将调查的所得数据统一上传到数据库/磁盘中，随后后台管理系统可以读取数据库的信息并进行相关处理，最后在前端页面上进行展示。其中，负责人可以利用后台管理系统根据人工审核结果对数据库的相关字段进行查询与修改。

本文设计的系统如图中虚线框内所示：需要对原有的调查问卷系统进行改进。将自动审核请求模块插入到原有的后台管理系统，自动审核请求模块会定时收集数据库中的调查问卷数据并向 API 接口发送请求，接口调用机器审核算法后，将结果返回，自动审核请求模块再将返回的结果存入数据库中。最后利用审核系统前端页面显示审核结果，并插入到原有系统的前端页面中，便于科普人员观看。

3.2 系统整体设计

为了实现调查问卷自动化审核系统的设计需求，本节将给出系统的整体设计，系统整体设计如图 3-3 所示：

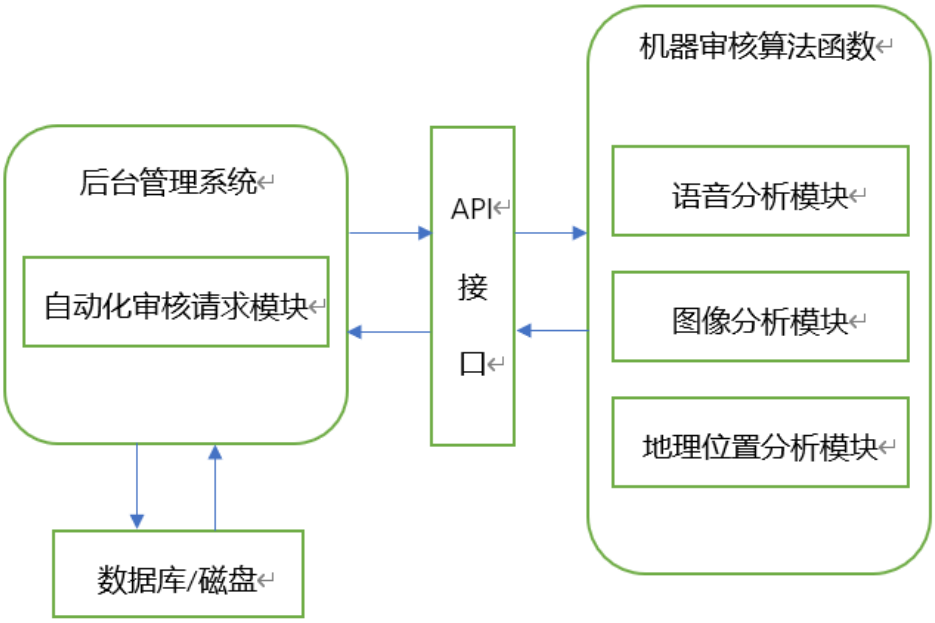


图 3-3 系统整体设计

Figure 3-3 Overall system design

从图 3-3 中可以看出，调查问卷自动化审核系统由三大部分组成：

（1）自动化审核请求模块：该模块部署在已有的后台管理系统中，需要自动将待审核的调查问卷从数据库/磁盘中读出，并向 API 接口发送审核请求，得到相关数据后存入数据库中。

（2）API 接口：由于后台管理系统与机器审核算法模块使用的编程语言不能直接互通，因此 API 接口主要进行两个模块间的信息交互工作，需要两个模块设计并遵守相应的规则。

（3）机器审核算法函数模块：该模块是本文系统的算法核心，主要包括三个部分：

1)语音分析模块：通过语音的识别及语义的分析判断调查问卷是否合格是本文最主要的审核算法。按照“科普调查规范”相关规定，通过语音识别后的文本与目标文本的相似度计算，可以定位错误的位置与错误的类型，解决了苏迪的方法中无法对不合格问卷进行解释的问题。

其难点在于受外界嘈杂环境及调查员、受访者的口音影响，语音识别出的文本可能与真实文本产生偏差。本文拟采用汉语拼音的音近词识别的方式来解决这一问题。

2)图像分析模块：该模块需要先对图像进行人脸检测，检测出图像中的人数，再截取图像中的人脸，进行性别的二分类识别，判断受访者的性别是否与调查问卷上所填信息一致。人脸识别算法采用了 MTCNN 网络，相比于苏迪方法使用的需

消耗大量算力的 YOLO 目标检测算法，MTCNN 算法使用的网络结构较为简单，在达到相同效果的前提下，只使用 CPU 便可以进行快速计算。提高了系统的运行效率。

3)地理位置分析模块：该模块计算了调查地点与目标居委会及区域内附近所有居委会的距离，根据距离之间的关系判断调查地点是否合乎要求。对比与苏迪方法中只设定阈值对调查地与目标居委会距离进行判断，本文的这种算法更加灵活，不容易造成误判。

在下面几节中，将依次介绍每个模块的结构设计。

3.3 API 接口连接原有系统与审核算法

本文设计的系统需要在科普机构原有的后台管理系统功能上，加入调查问卷自动化审核功能，因此需要连接原有后台管理系统与自动化审核算法。下面分别讨论设计的难点、解决方法，并给出设计方案。

3.3.1 设计难点与解决方法

该部分设计的难点是：需要在一个实际运营的系统上实现自动化审核功能。这需要做到原有系统与新增审核算法的连接。由于原有系统使用的编程语言是 PHP，而新增审核算法使用的编程语言是 Python，两种编程语言无法直接互通，因此无法将新增审核算法直接部署在原有系统中。另外，原有系统运行时已经消耗了原有服务器的大部分计算能力和内存空间，服务器无法承受更多的负载。

针对这个问题，本文采用了 API 接口的方法实现原有系统与审核算法的连接。新增加了一个服务器，在该服务器上开发了一个 RESTful API 接口，该接口可以调用审核算法函数并响应发送到该接口的请求。在原有服务器上设计脚本定时向接口发送请求并收集返回的数据，即可实现原有系统与审核算法的连接。由于审核算法部署在新增服务器中，因此不会加重对原有系统的负载。

3.3.2 接口设计

使用了 API 接口的自动化审核功能由运行在原有服务器上的自动化审核请求模块和运行在新增服务器上的审核接口模块构成，其流程如图 3-4 所示，图中左边部分部署在系统原有服务器中的自动化审核请求模块，右边部分部署在新增服务

器中的审核接口模块：

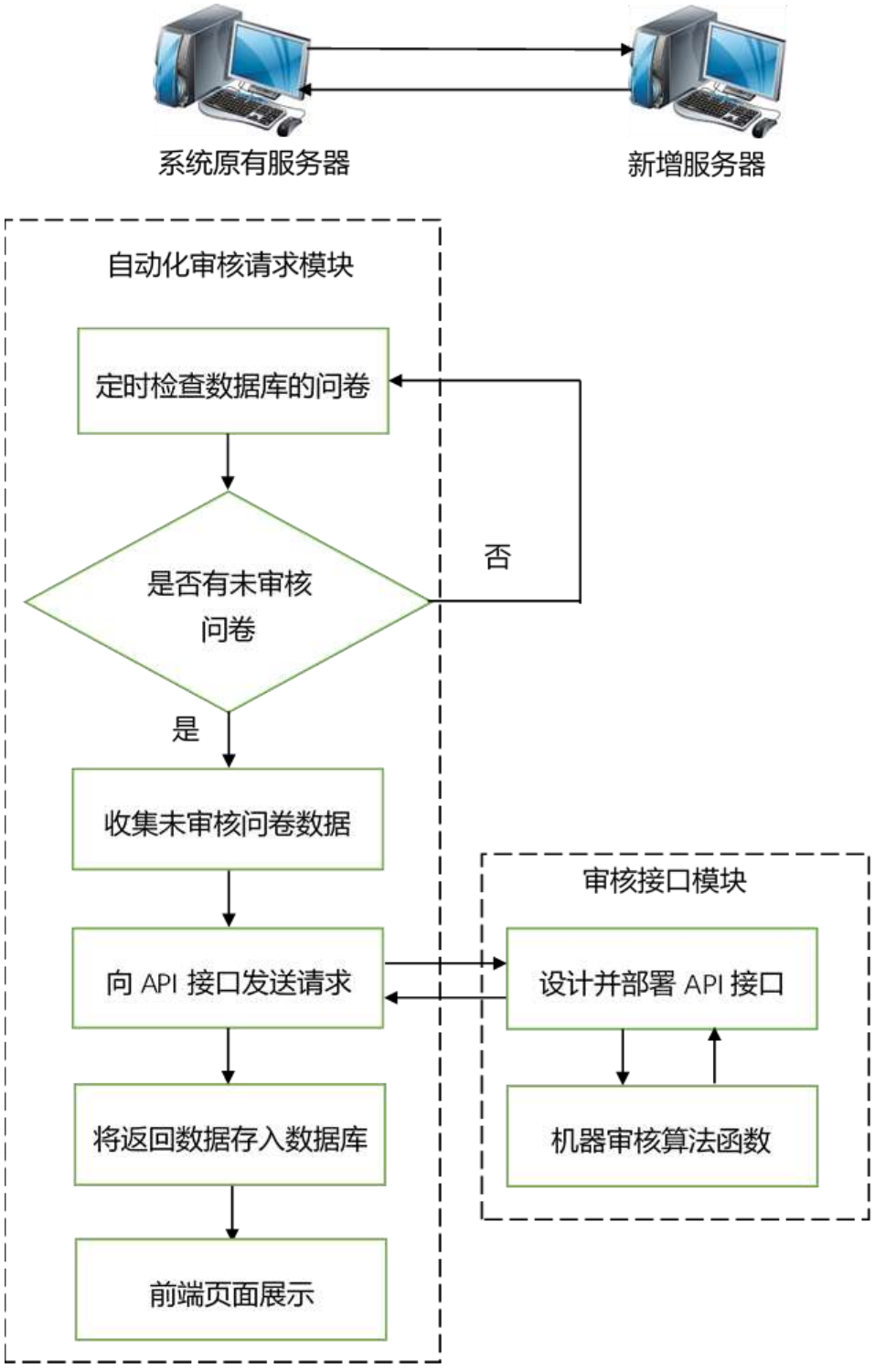


图 3-4 使用 API 接口的自动审核功能整体设计

Figure 3-4 Overall design of automatic audit function using API interface

由图可知，自动化审核请求模块首先需要在原有后台管理系统中对数据库中的问卷信息定时进行检查，如果没有尚未审核的问卷，则等待一个定时周期后再进行检查，若存在未审核的问卷，则需要从数据库及相应的磁盘中收集未审核问卷的

数据信息，并将这些收集到的数据进行打包，向 API 接口发送 HTTP POST 请求。部署在新增服务器上的 API 接口在接收到请求后，调用后续算法模块的相关函数对问卷进行审核，得到审核结果后，API 接口根据事先规定好的数据交换格式，将审核结果返回，后台管理系统将返回的数据保存到数据中，并设计程序进行前端页面的展示。

3.4 语音分析模块设计

为了使调查能够顺利进行，科普机构要求调查员在调查过程中必须遵守“科普调查规范”进行调查。该规范要求在调查问卷全过程录音中，调查员须完整的读出导语、结束语以及题目的题干、题项、选项等关键信息，不能够出现错读、漏读、不读等现象。

某些调查员为了节约时间成本，可能在朗读题目的过程中造假，例如：朗读时不念出具体的题目，而是用“第一题选什么”、“第二题选什么”等代替；部分调查员甚至在一旁打私人电话或是与他人聊天，让受访者自行填答问卷。这些作法违反了调查的要求，需要在审核时指出这些错误。

因此，本文引入了语音分析模块来解决上述问题。先将音频识别为对应的文本，再将这些文本与应该朗读出的题目进行文本相似度比较，判断调查员是否按要求进行朗读。

由图 3-5 可知，语音分析模块由五个部分组成：

(1)语音识别

考虑到语音识别需要非常高的准确性，且现有服务器配置的算力有限，无法在本地进行运算，因此本文的语音识别使用到了科大讯飞的语音转写技术，通过调用科大讯飞的语音转写接口，可以将调查问卷中的相关录音转换成相对应的文字，以便进行后续的分析。

(2)文本对齐题目区间

本文需要对调查问卷的各部分题目进行分析，所以需事先将音频转成的文本进行分段切片，并将其对应到具体的题目上。调查问卷的数据压缩包中有关于录音的日志文件，日志中详细标明了每道题目的开始时间以及结束时间。同样我们也可以从科大讯飞的语音转写接口中获得具体文本的起止时间，经过计算即可得出每道题目所对应的录音文本信息。

图 3-5 语音分析模块流程

Figure 3-5 Speech analysis module process

(3)文本预处理

文本预处理是自然语言处理中的关键步骤，也是后续研究的基础。本文中的文本预处理是对语音识别获得的文本和相关题目文本进行停用字符过滤及分词处理。停用字符过滤是值除去文本中所有标点符号及没有实际意义的词，本文按照算法需要编辑了停用字符表。文本分词是指将由字组成的文本以词为单位进行切分，本文使用了结巴分词工具对语音识别后的文本以及题目文本进行了文本分词处理。

(4)基于汉语拼音的目标文本修正

由于调查过程中可能出现噪音干扰、调查员发音不准以及词语同音不同义的问题，语音识别有时并不能准确的识别录音所述内容，会出现一定的偏差，例如：“科普”识别为“客服”、“中专”识别成“终端”、“进展”识别成“近战”等。因此本文利用汉语拼音的结构特点，将每个汉字拆分成声母和韵母两部分并进行模糊识别，绝大多数两个字词语识别错误的问题中，汉语拼音只相差了一个声母或是一个韵母，有的只是同音词声母韵母完全相同。本文根据题目中的目标文本使用汉语拼音模糊识别的方式对语音识别后的文本进行了修正，大大提高了语音识别的准确性。

(5)文本相似度计算

本文针对“科普调查规范”要求进行分析，总结出了每道题目中需要出现的文本，利用上述步骤得到的题目对应的修正后的录音文本信息，使用编辑距离法、Jaccard 文本相似度法以及 Word2Vec+余弦距离的方法对它们进行了文本相似度计算，得到题目中每个小部分所对应的相似度得分。科普机构可以根据得分设定阈值自动判断问卷可信度，也可以通过得分辅助审核员进行判断。

这个模块的难点在于，由于背景中存在噪声或是调查员发音不准的问题，对音频的语音识别可能会产生偏差。例如：将“科普”识别为“客服”，“中专”识别为“终端”等，这些语音识别上产生的错误会致使算法在计算文本相似度时的结果偏小，在阈值判定时被判定为不合格。在这种情况下，调查员实际上正确读出了需要朗读的题目，却因为语音识别偏差导致了误判。

本文使用了基于汉语拼音的修正方法解决了这一难点。将汉字以汉语拼音的方式拆分成声母和韵母，如“科普”将“科”拆分为声母“k”与韵母“e”，“谱”拆分为声母“p”与韵母“u”，与“客服”拆分成的“k”、“e”、“f”、“u”，“科普”与“客服”的四个声韵母中，只相差了一个声母，因此在相似度分析时可以将识别出的“客服”修正为题目中应出现的“科普”，从而提高后面文本相似度的计算结果，防止误判的发生。

3.5 图像分析模块设计

问卷调查过程中会对调查员和受访者进行随机拍照，“科普调查规范”规定，调查需要发生在调查员与受访者一对一的环境中，因此照片中需要出现调查员和受访者的人脸图像，并且受访者的性别需要与调查任务中的目标性别一致。

在以往的调查中，有些调查员为了提高调查的效率、谋取更多收益，会做出以下违规行为：(1)前往人数较多的公共场所进行调查（非一对一环境）(2)使用变声器假扮受访者，自问自答（没有受访者）(3)为了完成性别任务指标，对受访者的性别信息进行修改（性别信息造假）

虽然这种问题出现情况较少，但同样需要引起重视。由于现有的语音识别技术无法准确识别出背景环境/变声器/发音人性别，因此需要借助图像分析模块进行分析。通过对照片进行人脸检测，查看照片中是否存在两人，判断调查环境是否为调查员与受访者一对一的环境。通过对照片中出现人物的性别检测，判断受访者真实性别是否与所填信息一致。

图像分析模块的流程如下：

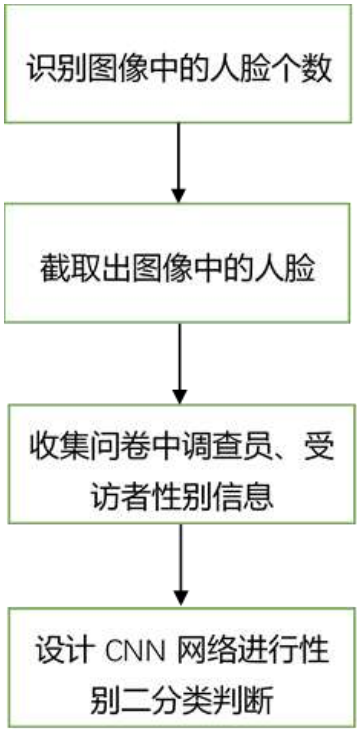


图 3-6 图像分析模块流程

Figure 3-6 Image analysis module process

由图可知，首先需要根据调查问卷中的图像数据，使用 MTCNN 人脸检测方法检测出图中人脸的个数，所有图片中至少需要有一张图片检测出两张人脸。然后对这两张人脸进行截取并将图片调整为合适的大小，从调查问卷的数据中获取调查员以及受访者的性别，如：“一男一女”、“两男”、“两女”。最后对截取的人脸分别进行人脸二分类检测，判断性别统计是否与调查问卷数据中的记录相一致。

本文通过对调查问卷中的实际样本进行人脸截取和性别打标，作为真实数据训练出了卷积神经网络（CNN）记性了性别二分类判断。

3.6 地理位置分析模块设计

科普机构给问卷调查制定了任务，为保证调查结果的科学性，调查员需要根据任务需求前往指定的社区/居委会/村委会进行调查任务。

然而，由于部分地区的居委会调查难度较大、位置较偏、亦或是调查员在完成某地任务已前往下一地点的情况下又被额外增加了任务，受利益驱使，部分调查员就近选了居委会而并没有前往指定的居委会进行调查。例如：调查员的任务时前往居委会 A 进行调查，然而居委会 A 实行了封闭管理，需要办理相关的手续方可进入调查，调查员不愿意办理手续，遂前往附近的居委会 B 进行调查，并将调查信

息打上居委会 A 的标签。

这严重影响到了调查结果的可靠性，因此本文针对调查的地理位置进行了分析。先计算调查地点与目标居委会的距离 d1 以及调查地点与附近其余所有居委会距离的最小值 d2，对 d1 与 d2 进行分析，设置阈值判断地理位置是否合格。

针对地理位置进行分析需要借助全球定位系统（GPS）来完成，GPS 是基于卫星定位的系统，它可以计算出使用者在地球表面具体位置的经纬度坐标，人们可以根据两点的 GPS 坐标来计算两点间的距离。计算距离的公式如下：

$$d = 2 \cdot arcsin\sqrt{\sin^2\left(\frac{\delta_1-\delta_2}{2}\right) + cos\delta_1 \cdot cos\delta_2 \cdot sin^2\left(\frac{\alpha_1-\alpha_2}{2}\right)} \quad (3-1)$$

其中，两点经纬度坐标分别为(α_1, δ_1)和(α_2, δ_2)，d 为两点之间的距离。

地理位置模块的设计如图 3-7 所示：

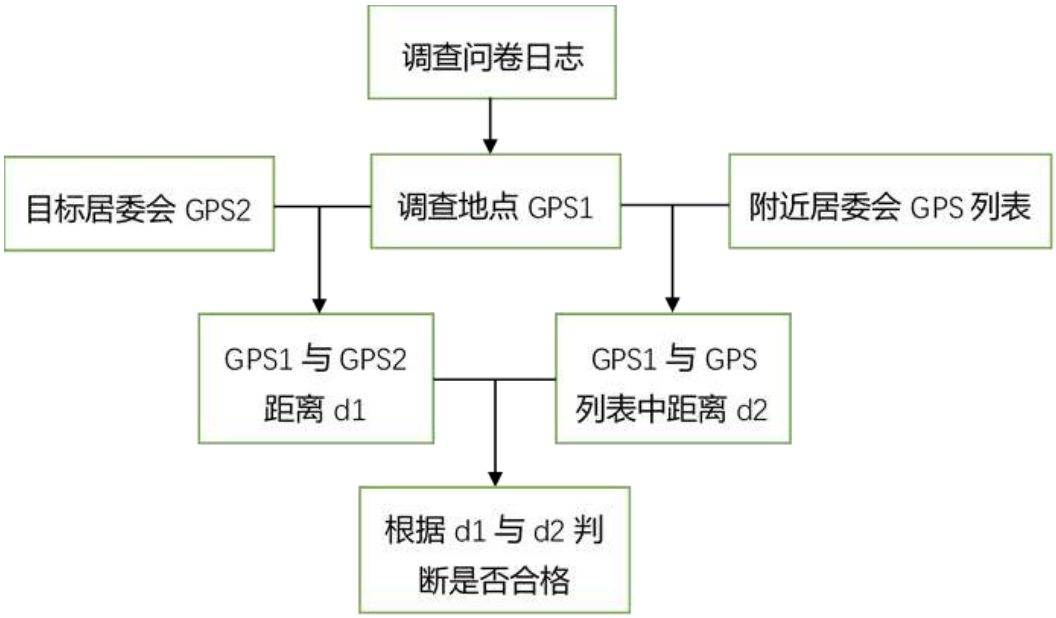


图 3-7 地理位置分析模块流程

Figure 3-7 Geographical analysis module process

由图 3-7 可知：调查问卷 APP 在进行调查时，设备收集了调查时的 GPS 信息，并存入日志文件中。我们从调查问卷的日志中获取调查地点的信息 GPS1，根据问卷记录的目标居委会名称查询到目标居委会的信息 GPS2，计算出调查地点距目标居委会的距离 d1。根据目标居委会名称，我们可以获得附近所有居委会的 GPS 列表，计算所有的距离并找出最近的居委会距离 d2。根据 d1 与 d2 进行分析，判断问卷是否合格。

一般来说，调查地点应该距离目标居委会最近，因此 d1 的距离应该小于 d2，若 d2 距离大于 d1 且数值较大，则极大可能为不合格问卷。

如果在调查问卷日志中没有记录调查地点的 GPS 信息，则按照不合格问卷进行处理。

3.7 系统性能优化

3.7.1 音频去静默

在调查进行的过程中，受访者听到问题以后会因对问题的思考而陷入沉默，调查员在调查时也会因为对调查问卷 APP 的操作而短时间沉默，而调查问卷的录音功能会记录从开始调查一直持续到调查结束的所有音频，所以调查问卷的录音中会存在较长一部分的静默片段。由于这些静默片段本身不含有任何有用的信息，而本文系统进行语音识别时需要调用科大讯飞的语音转写接口，接口转写所需要的时间和业务量计算都是由音频时长决定的，因此这些静默片段的存在会增加系统的时间和财力成本，造成不必要的浪费。

本文所设计的系统将在语音分析模块进行语音识别之前，增加一个音频去除静默的步骤，利用音频的功率大小检测出调查问卷录音中不含有可用信息的静默部分，将其去除之后再进行语音识别。这样做可以减小系统运行的开销，提升系统运行的效率。

3.7.2 文本检测效率优化

虽然调查问卷日志文件已经标注了 A1-D5 中所有大题题目的区间，通过这些区间的音频获得对应的录音文本，这些文本可以与该题目中的题干、题项、选项等进行文本相似度检测，但是日志中标注的区间所含有的文本数量依旧十分庞大，检测所需时间较长。

本文所设计的系统，通过对问卷中题目与调查问卷录音文本的分析，发现大题题目区间内的各个题干、题项、选项的存在着顺序关系，例如：调查员一般首先朗读题干，然后按照题项的顺序依次朗读出题项，而选项一般会存在于题干或某一题项之后。因此我们可以按照题干>>题项 1>>题项 2>>……>>题项 n 的顺序进行检测。首先检测区间内所有内容检测到题干，然后用题干后的内容检测题项 1，用题项 1 后的内容检测题项 2，以此类推，最后我们利用题干后的所有内容检测选项。这样做进一步缩小检测的区间，避免了每次对所有区间进行检测，减少了文本相似度检测所需要的时间，提高了系统的运行效率。

3.8 本章小结

本章首先对本文研究的调查问卷审核问题进行了分析，然后给出了调查问卷

自动化审核系统的总体设计，接着设计了连接原有系统与本文的审核算法的 API 接口的整体流程，之后分别设计了用语音识别、文本相似度分析技术的音频分析模块，使用 MTCNN 人脸检测、CNN 性别二分类识别实现图像分析模块，设计了地理位置分析模块，这个模块使用 GPS 计算距离并加以分析问卷的合格性，最后给出了两种系统性能优化的方法。

4 相关模块具体设计

本文的第三章给出了系统的各部分的整体设计思路与大致的实现方法。本章将详细论述系统各部分的具体设计细节。共分为 5 个小节，第一节给出了使用 API 接口连接原有系统与本文审核算法的具体实施细节，第二节重点给出了使用语音识别、静默检测及文本相似度检测的语音分析模块具体方法作为本文的算法核心，第三节给出了使用 MTCNN 人脸检测识别调查环境以及性别二分类 CNN 网络的设计，第四节给出了使用 GPS 进行地理位置分析和审核的方法，第五节为本章小结。

4.1 API 接口连接原有系统与审核算法

本文设计使用 API 接口的方法连接原有的系统与审核算法，本节具体介绍了图 3-4 中原有服务器中的自动化审核请求模块以及新增服务器中的 API 接口具体设计与部署。

4.1.1 新增服务器上的 API 接口设计

根据本文系统需要，原有后台管理系统需要主动与审核算法模块进行通信，向审核算法模块提供调查问卷的数据，并且保存得到的返回结果。所以我们需要在审核算法侧部署一个 API 接口，可以对原有后台管理系统提出的请求进行响应。

本文采用 Flask 作为 API 接口的编程框架。审核算法使用的语言是 Python，所以我们需要使用 Python 的 Web Server 来部署一个 API 接口。目前 Python 中主流的 Web Server 框架是 Django 与 Flask。由于本文中审核算法侧的设计只需要满足简单的接口功能，涉及到的业务较为简单，因此本文不选择使用 Django 这种重量级的框架，而是选择了 Flask 这种自由、灵活的轻量级框架作为开发工具。使用 Flask 编写 API 接口，不需要安装许多额外的插件或组件，也不像 Django 存在很多规则和约束，因此部署时不太需要复杂的操作，代码也会相对简洁，非常适合本文设计的系统。

有关接口设计具体流程的 Python 代码如图 4-1 所示：

```python
import zipfile
from flask import Flask, jsonify, request
from flask_script import Manager
from function import questionnaire_verify, del_file, IP_LIST
app = Flask(__name__)
manager = Manager(app)  # 启动Web Server
app.config['JSON_AS_ASCII'] = False  # 配置文件
app.config.update(DEBUG=True)

@app.route('/', methods=['POST'])  # 收到post请求，上传文件到本地
def api_upload():
    if request.remote_addr not in IP_LIST:  # 验证IP地址
        return jsonify({"error": "非服务器ip"})

    f = request.files['file']  # 从表单的file字段获取文件，file为该表单的name值
    azip = zipfile.ZipFile(f)
    fname = f.filename.split('.')[0]  # 调查问卷名称
    for file in azip.namelist():  # 文件解压到相应位置
        azip.extract(file, r'./download/' + fname)
    azip.close()

    result_dict = questionnaire_verify(r'./download/' + fname)  # 调用审核函数
    del_file(r'./download/' + fname)  # 删除相关文件

    return jsonify(result_dict)  # 返回JSON格式结果

if __name__ == '__main__':
    app.run(debug=True, host='0.0.0.0', port=10000)
    # manager.run()  # 命令行启动方式
```

图 4-1 API 接口实现代码

Figure 4-1 API interface implementation code

由图 4-1 可知，接口设计主要包括设置 HTTP 请求格式、验证 IP 地址、接收数据并传入函数、返回结果四个部分：

(1)设置 HTTP 请求格式

第 10、28 行设置了 HTTP 请求格式。API 接口使用的应用层通信协议是 HTTP 协议，该协议包含了 GET、POST、PUT、DELETE 四种方法。由于该接口需要接收原有后台管理系统发送的数据并作出响应，因此使用了 POST 请求方式。具体方法为在接口配置“@app.route”中加入“methods=[‘POST’]”，并将“app.run”函数中的“port”值设置为 10000、“host”值设为“0.0.0.0”（对应当前设备 IP 地址）。这样后台管理系统程序就可以使用 POST 指令根据目标 IP 地址发送 URL 请求到服务器的 10000 号端口进行接口的调用。

(2)验证 IP 地址

第 12、13 行用于验证 IP 地址。出于对系统安全性原则的考虑，我们不允许后台管理系统之外的程序访问该接口，以避免数据的泄露。在接口正式处理请求之前，需要先加入一小段程序来检验 URL 请求的发送 IP 地址是否为后台管理系统服务器所对应的 IP，我们首先查看一下原有后台管理系统所使用的服务器 IP 地址，然后在接口中加入 Python Flask 中的“request.remote_addr”指令来查询向接口发送请求的 IP 地址，与原有服务器 IP 地址作对比，过滤掉来自非后台管理系统服务器的

IP 地址的请求，保证系统的安全。

(3)接收数据并传入函数

第 15 到 23 行用于接收数据并传入函数。由于算法模块需要得到调查问卷的相关信息并进行分析，因此接口需要使用“request.files”接收通过表单传过来的数据流，并通过事先约定好的“file”字段获得算法模块需要的问卷包压缩文件流（格式为 ZIP 压缩格式），使用 zipfile 函数解压到本地进行存储，并将存储位置告知审核算法模块函数，调用函数“questionnaire_verify”进行分析。为了对调查数据进行保密，同时防止大量问卷堆积在磁盘中导致系统运行效率下降，在函数执行完毕后，需要使用“os.remove”对存储在本地有关调查问卷的信息进行删除。

(4)接口返回 JSON 结果

第 25 行用于返回 JSON 结果。审核算法模块函数对调查问卷数据处理后会生成相应的字典格式的数据，由于 Python 中的字典（Dict）功能相当于 PHP 中的数组（Array），但两者的格式类型不同，无法直接进行转换，因此接口需要通过专用的 JSON 格式进行交互。在接口返回结果时使用 Flask 中的“jsonify”函数先将字典变量转换成 JSON 格式，再将结果返回给请求方。

4.1.2 API 接口的部署

图 3-4 中的原有服务器在实际应用中是租用云服务器实现的，本文中新增的服务器在实际系统中也是租用云服务器实现的。

因此在本地环境下将接口测试完毕后，需要将接口部署到新增的用于实现审核算法的云服务器（图 3-4 中的右侧）上以满足实际生产中的需要。为了不给原有后台管理系统的服务器增加负担，本文使用了新增拥有双核 CPU 和 4G 内存的阿里云轻量级应用服务器。在本地使用 Xmanager 的 Xshell 终端模拟软件远程登录到云服务器上对接口进行部署。

使用 SSH 方式远程登录到 Linux 服务器上，利用 pip install+所需包名（flask 等）的指令使得云服务器上的 Python 环境与本地一致，再使用 Xmanager 中的 Xftp 软件将写好的 Python 脚本拷贝到服务器对应的文件夹下。由于接口使用到 10000 号端口，需要开放该端口，因此在服务器的/etc/sysconfig/iptables 文件中加入代码“-A INPUT -p tcp -m tcp --dport 10000 -j ACCEPT”来解除防火墙的限制。

由于我们使用了 Xshell 软件对终端进行操作，因此当我们关闭 Xshell 软件断开 SSH 连接时，系统会自动中断所有运行中的脚本程序。我们希望程序一直处于运行状态，因此需要借助“nohup”与“&”指令。其中，“nohup”指令代表程序将永久的执行下去，不会因用户断开连接而中断，“&”指令代表将脚本任务放到后

台进行运行。使用“nohup 脚本指令&”的方式，我们就可以将接口部署在服务器，且在后台一直运行下去，有关接口的相关信息会被保存在同目录的“nohup.out”文件中便于后续查看。

在执行接口的脚本文件是有两种方式，一种是直接运行接口脚本的单线程方式，这种情况下接口只能同时处理一个请求，其余的请求会在当前请求结束后才进行处理，这种接口效率较低。第二种方式是通过指令，开启接口的多线程模式，这种能够情况下接口可以同时并行处理多个请求，充分利用系统资源，提升了运行效率。最终在终端中输入的指令如下：

“nohup python3 app.py runserver -h 0.0.0.0 -p 10000 --thread &”

这样就完成了并行的接口在服务器上的部署，后台管理系统可以通过对服务器发送 HTTP 请求来调用接口并使用接口内的审核算法函数进行自动化审核。

4.1.3 原有服务器上的自动化审核请求模块

本小节实现图 3-4 中的左侧的自动审核模块的流程图。对原有的后台管理系统进行改进，将每一份问卷调用接口进行审核算法的计算。该模块实现的基本功能是周期性的查看是否有未审核的调查问卷，如过有则调用脚本进行一次批量审核，如果没有则继续等待直到新问卷的出现。

(1)审核任务实现脚本

原有的后台管理系统使用的是 PHP 的 yii2 框架，因此脚本文件也沿用了这一框架，并继承了“CConsoleCommand”类，用于后面执行定时任务。脚本首先会使用“Yii:app()->db()”指令连接数据库，查看数据库中存放问卷信息的“lime_survey_check”字段（0 值为尚未进行自动化审核，1 值为已完成自动化审核），并获取前 10 个字段值为 0 的调查问卷编号。然后从数据库的“de_surveypath”表中找出这 10 个问卷的压缩文件数据在磁盘中存放的位置。接着用 PHP 中的 cURL 函数将这些压缩文件使用 HTTP 的 POST 请求传给 API 接口，在得到接口的响应后，将对应问卷“lime_survey_check”字段的值从 0 改为 1，代表该问卷审核完成，并将响应的结果存储到数据库的“de_survey_check”表中，以便后续的查询。

(2)周期性调用审核任务

由于科普调查进行时，会有源源不断的问卷信息被上传到数据库与磁盘中，而每次脚本只能够处理 10 份问卷，因此需要每隔一段时间执行一次该脚本才能够保证所有问卷都能进行审核。文本采用 Linux 系统中的 crontab 功能，定时执行脚本。

然而在一个定时周期内，接口不一定能够完成该脚本的所有请求，此时如果再次执行脚本，会导致后续的请求堆积在接口上造成拥塞，可能会导致系统的崩溃，

因此在执行定时任务时需要使用 flock 对任务加锁，使得时间到达定时周期后，若前一次脚本没有执行完毕（有请求未获得响应），则需等待下一周期的来临才能执行新的任务。

具体做法为：在/var/spool/cron/root 文件中加入指令：

“5,20,35,50 **** flock -xn /temp/conti.lock php AskCheck.php >> AskCheck.log”

这段指令代表，每个小时的第 5、20、35、50 分钟（即每隔 15 分钟），将运行一次 php 的 AskCheck.php 脚本，并产生一个 conti.lock 的锁文件，当脚本结束时会将锁文件删除。一个周期到来时，若锁文件存在，则本次不执行脚本，防止冲突的发生，最后将脚本执行结果存入 AskCheck.log 文件中。

4.2 语音分析模块具体设计

语音分析模块是调查问卷自动化审核系统的核心模块，本节给出 3.4 节中语音分析模块的具体实现，依次介绍了使用静默检测提取有效音频，调用接口进行语音转写与识别，区间对齐、分词、文本修正等预处理工作，用于审核的文本相似度计算以及缩小检测区间提高系统性能的方法。

4.2.1 静默检测提取有效音频

在进行调查的过程中，APP 会记录从调查开始到调查结束全过程的录音信息，由于调查过程中，受访者可能会因为对问题的思考而陷入短时间沉默，调查员也可能会因为对调查问卷 APP 进行的切题等操作而陷入沉默，因此在音频中会出现一些静默片段，这些静默片段并没有实际的意义，需要在语音分析模块开始时检测并去除这些片段，获得其中含有人声信息的有效音频。

Python 中用于静默检测的工具是 pyhub 库，先使用“AudioSegment.from_file”函数读取音频文件，将音频文件的每一帧转成 pyhub 可识别的 float 格式并存储在一个列表中。对列表使用“detect_silence”函数进行静默检测，截取出其中的非静默片段，该函数的可选参数有：最小静默检测时长“min_silence_len”、静默检测阈值“silence_thresh”、静默检测步长“seek_step”。

“detect_silence”函数首先将以分贝（db）为单位的“silence_thresh”转化为 pyhub 中标准 float 格式得到阈值，然后从音频的开头起步，每隔一个“seek_step”长度截取一段长度为“min_silence_len”的片段，并对该片段进行 RMS（均方根值）计算，RMS 公式如下：

$$X_{rms} = \sqrt{\frac{\sum_{i=1}^{N} X_i^2}{N}} = \sqrt{\frac{X_1^2 + X_2^2 + \cdots + X_N^2}{N}} \quad (4-1)$$

从公式中可以看出，RMS 公式与功率计算的公式十分类似，因此 pyhub 使用的静默检测原理实际上就是对声音信号的平均功率进行检测。虽然这个方法并不是专门针对人声进行的检测，但是由于在大多数问卷中，音频主要为调查员和受访者两个人的对话声音，因此该方法可以胜任本文系统中静默检测的任务。

随后 Pyhub 将计算得到 RMS 结果与转化后的阈值比较，如果 RMS 结果较小，则该段声音被检测为静默音频，在对所有音频进行检测后去除这些静默的音频。最终获得了含有人声信息的有效音频，用于后续的操作。

4.2.2 语音转写接口的调用

本文算法需要将调查问卷中的录音信息转换成相应的文本，并对文本进行识别，因此需要对通过静默检测后的有效录音进行语音识别。

文本语音识别采用的方法为调用科大讯飞的语音转写功能，该功能有 API 和 SDK 两种实现方式。由于 SDK 需要的运行环境为 Java 编程环境，因此本文采用调用 API 接口的方式进行语音识别。

(1)格式转换

科大讯飞的语音转写接口需要输入的音频格式为 wav/mp3 格式，而调查问卷中的录音信息存储格式为 aac 格式，所以在正式进行接口调用之前，需要对音频编码格式尽心转换。使用到的工具是音频处理软件 FFmpeg 的 Python 版本，对应的代码为：

“ff = FFmpeg(inputs={音频名称+'.'+'aac': None}, outputs={音频名称+'.'+‘wav’: None}) ff.run()”

即可将 acc 格式的音频转成 wav 格式，作为后续调用接口时的输入数据。

(2)接口调用

本文使用 Python 中的 requests.post 方法调语音转写接口，需要输入的数据为“data”字段中设置的参数、HTTP 请求的 url 地址以及“files”字段中对应的音频文件。

接口的“data”字段设置的部分相关参数如下表：

表 4-1 语音转写接口相关参数

Table 4-1 Related parameters of the voice transliteration interface

参数名称	参数类型	参数说明
app_id	string	讯飞开放平台应用 ID
task_id	string	转写任务的 ID
signa	string	加密数字签名
file_name	string	文件名称
file_len	string	文件大小
speaker_number	int	发音人个数
has_seperate	string	转写结果是否包含发言信息
language	string	cn:中文（默认）；en：英文

其中 app_id 是由科大讯飞官网上进行申请，并在开启服务后获得对应的 secret_key，使用 Python 中的 hashlib 和 hmac 包对 app_id 和 secret_key 进行加密，得到数字签名 signa。将调查问卷的 ZIP 数据包中的音频文件加入 POST 请求的 file 中，并将其名称和长度设置到参数里。设置发音人个数为 2（即调查者和受访者），has_seperate 为 True，language 为默认的 cn。

由于科大讯飞的服务器 IP 并不固定，因此采用域名作为 url 地址的方式进行调用，域名为'http://raasr.xfyun.cn/api'+请求的服务名称，其中'/prepare'用于生成转写任务的 taskid 并验证数字签名，'/upload'为上传音频，'/getProgress'为获取任务进度，'/getResult'为获取语音识别结果。upload 上传音频后，每 20s 发送一次 getProgress 请求查看任务进度，如果任务完成则使用 getResult 获得任务结果，并将该结果以 JSON 的格式保存到磁盘中供后续算法程序调用。

(3)返回数据

任务完成后，接口的部分 JSON 结果如图 4-2 所示：

图 4-2 接口返回的 JSON 数据

Figure 4-2 JSON data returned by the interface

图 4-2 中，bg 字段对应的值为句子相对于音频的起始时间，单位为 ms；ed 字段对应的值为句子相对音频的终止时间，单位为 ms；onebest 字段为句子的内容，speaker 字段为说话人编号，从 1 号开始。

4.2.3 预处理工作

(1)文本与题目区间的对齐

为了缩小后续进行文本相似度检测的区间，减轻系统运算时的压力，需要计算出语音转写后的文本所对应的调查问卷的题目编号与区间。

调查问卷的日志文件中记录了录音的开始时间以及问卷中每个题目的开始时间，日志文件的记录如图 4-3 所示：

```
2019-05-15 14-26-49 SSS,recorder:start:fileName:rec-1.aac&&
2019-05-15 14-27-12 SSS,recorder:welcome_start_button&&
2019-05-15 14-27-12 SSS,recorder:stop_welcome&&
2019-05-15 14-27-12 SSS,wifi mac:F49FF36214A1&&
2019-05-15 14-27-12 SSS,survey:start&&
2019-05-15 14-27-13 SSS,recorder:start&&
2019-05-15 14-27-25 SSS,survey:load&&
2019-05-15 14-27-30 SSS,question:A3:A3.您的户籍是否在<font color="#FF0000">本市</font>？ &&
2019-05-15 14-27-34 SSS,question:A4:A4.民族&&
2019-05-15 14-27-42 SSS,question:A5:A5.您现有的文化程度？ &&
2019-05-15 14-27-55 SSS,question:A5a:A5a.您在读高中（中专、技校）时，所学的课程是偏文科还是偏理科？ &&
2019-05-15 14-28-24 SSS,question:A5b:A5b.到目前为止，您总共受过几年的正规学校教育？ &&
```

图 4-3 日志文件记录

Figure 4-3 Log file record

由图中信息可以获知每道题目对应的音频区间，例如：录音开始于 2019 年 5 月 15 日的 14 时 26 分 49 秒，题目 A3“您的户籍是否在本市”开始于 27 分 30 秒，A3 的下一题 A4 开始于 27 分 34 秒，因此可以得出，题目 A3 在音频中对应的区间大概在第 41 秒与 45 秒之间。由于日志的记录只精确到了秒，因此为了保证区间的完整性，我们会将 A3 题目的区间向外扩大一秒，即 40 秒到 46 秒之间。

语音转写接口返回的结果中通过“bg”和“ed”字段给出了转写后的文本出现在录音中的时间位置，与上述从日志文件分析出的题目对应的区间对比，就可以计算出每个题目所对应的录音中的文本。

这里需要注意其中的一些细节，由于语音转写时使用的音频文件是经过静默检测提取后录音的有效部分，并非原始音频，因此我们需要对静默检测时被去除的部分进行记录，并通过这些记录对语音转写后的“bg”与“ed”对应的时间区间进行“还原”，防止时间轴发生错乱。

使用这种做法可以将相似度检测的区间从全部文本缩小到了一个区间中，大大减小了检测时计算量。

(2)停用字符过滤与分词

停用字符过滤与分词都是自然语言处理中对文本进行预处理的关键步骤。在进行文本相似度分析前，我们需要将文本中没有意义的字符进行过滤，例如标点符号与汉字中的语气词、连接词，如“的”、“呢”、“吗”、“嗯”等词语。

分词就是识别出词语的边界，在由多个汉字组成的文本中以词为单位进行分割。目前已有很多成熟的汉语分词技术，本文使用了速度较快、效果较好而且可以在 Python 上进行使用的结巴分词工具。

结巴分词可以使用精确模式将句子切成多个词语，调用函数的方法为：

“words_list = jieba.cut(sentence,cut_all=False)”

例如，对“在过去一年中，您是否去过下列科普场所？如果去过，主要是因为哪个原因？”这个题目进行停用字符过滤及分词，分词结果为【在、过去、一年、中、您、是否、去过、下列、科普、场所、如果、去过、主要、是、因为、哪个、原因】。本文中主要使用结巴分词辅助汉语拼音对文本进行修正、以及将文本拆分成元素进行文本相似度分析。

(3)基于汉语拼音的文本修正

受到调查员发音不准、背景有杂音等问题的影响，语音识别出来的文本结果往往会与实际调查中的真实情况不符。本文根据对多份问卷的语音转写文本进行分析，发现这类错误往往出现在发音比较相近或是同音不同字的两字词语上。

因此，本文计划使用汉语拼音的方法根据题目中的目标文本对语音识别出的文本进行修正。首先，利用结巴分词的精确模式将题目中的目标文本（如题干、题项、选项等）分成一组词语，将其中由两个字组成的词语保存在一个目标列表中。再将题目对应的语音识别出的文本以长度为 2，步数为 1 的方式（如将“我是研究生”截取为“我是”、“是研”、“研究”、“究生”），截取出所有的两字词语，最后将目标列表中的词语与这些截取出的词语作对比，如果满足相似的条件，则使用目标列表中的词替换文本相应位置的词语。需要满足的条件如下：

1)将词语中的汉字按照汉语拼音的形式分为声母与韵母，两个词语各包含了两个声母与两个韵母共四个部分，如果两个词语的四个部分中，有三个或四个部分相同，且两个词语本身不同的情况下，可以认定为相似。比如“中专 zhongzhuan”与“终端 zhongduan”之间只差了一个声母“d”与“zh”，满足条件。

2)在对 1)的实践中发现，由于目标文本的词语中本身就存在着互为相似词语的情况，因此相互间的替换反而导致修正效果适得其反（例如，目标文本与待修正文本均为“ABCD”，其中 A 与 D 互为相似词语，则按照 1 中条件可能会修正为

“DBCA”），因此在替换时还需要满足，若截取出的被认定为相似的词语本身就存在于目标列表中，则不进行替换。

本文中对汉字的声母、韵母分解使用了 Python 中的 pypinyin，调用“lazy_pinyin”函数实现，方法为：

“f1 = lazy_pinyin(w1, style=3)

f2 = lazy_pinyin(w1, style=5)”

其中，“style”为 3 代表输出声母，为 5 代表输出韵母。

利用汉语拼音对文本进行修正可以适当解决语音识别不准的问题，提高后续计算出的文本相似度，防止对部分正确问卷的误判。

4.2.4 文本相似度计算

“科普调查规范”中要求调查员读出调查问卷题目中的目标文本，本文通过计算修正过的录音文本与题目中目标文本的相似度，并设定阈值来检测调查员是否按照规范进行调查。

对调查问卷中的目标文本的检测主要存在于问卷的 B 部分（公民的科技信息来源）、C 部分（公民对科学的理解）、D 部分（公民对科学技术的态度），每部分含有若干大题。本文需要对大题中的题干、题项和选项进行文本相似度检测。B2 题目如图 4-4 所示：

B2. 您对下列**科技发展信息**是否感兴趣？感兴趣的程度如何？（每行选一项）

题 项	非常感兴趣	一般感兴趣	不感兴趣	不知道
(1) 宇宙与空间探索	○	○	○	○
(2) 环境污染及治理	○	○	○	○
(3) 计算机与网络技术	○	○	○	○
(4) 遗传学与转基因技术	○	○	○	○
(5) 纳米技术与新材料	○	○	○	○
(6) 新能源开发及利用	○	○	○	○

图 4-4 公民科学素质抽样调查问卷 B2 题

Figure 4-4 Sampling questionnaire for citizen scientific quality Question-B2

其中，“您对下列科技发展信息是否感兴趣？感兴趣的程度如何？”为题干，“非常感兴趣、一般感兴趣、不感兴趣”为选项，“宇宙与空间探索、环境污染及治理、计算机与网络技术、遗传学与转基因技术、纳米技术与新材料、新能源开发及利用”为题项 1 到题项 6。

我们需要从科普机构提供的问卷问题信息明细中将 B、C、D 部分中每道大题的题干、题项、和选项的文本提取出来，并以字典的形式保存在 Python 可以识别

的“.pkl”文件中，例如“题项”保存为“B2-1：宇宙与空间探索”、“选项”保存为“B2-0：非常感兴趣、一般感兴趣、不感兴趣”、“题干”保存为“B2：您对下列科技发展信息是否感兴趣？感兴趣的程度如何？”等。

随后我们需要读取“.pkl”文件中的字典作为目标文本，针对 4.2.2 节中题目所对应的文本区间进行文本相似度检测，具体检测过程如下：

(1)首先根据字典键值对中的“键名”找出对应的答题题目，如键名“B2-1”对应的题目“B2”。

(2)再找出题目所对应的录音区间使用语音转写获得的文本，对该段文本以及字典对应的值（目标文本）进行停用字符过滤操作，如“宇宙空间与探索”。

(3)计算停用词过滤后的题目的目标文本长度，取值为 N，设置一个窗口值为 L（L 取值为 N 或 N+1）的窗口对语音转写获得的文本进行截取，步长取值为 S（通常为 1）。这样就可以获得录音文本中的若干长度为 L 的文本片段（片段数量由步长 S 决定）。

(4)利用文本相似度计算的方法，计算题目目标文本与所有截取出的片段文本之间的相似度值，取其中的最大值作为该题目与录音文本的最终相似度值，即调查员念出对应题目的得分，设置阈值对得分进行判断，判定调查员是否按照规范进行调查。

本文进行文本相似度计算的方法有三种，分别为编辑距离算法、Jaccard 相似度算法、以及 Word2Vec 算法。三种方法分别使用到了 Python“distance”库中的“levenshtein”函数、“sklearn”库中的“CountVectorizer”函数以及加载了 Word2Vec 模型的“gensim”库，可以计算出相应的相似度得分。

使用编辑距离算法计算时，记录两字符串的编辑距离为 a，再除以目标文本长度 N，得到 a/N，将 1-a/N 作为编辑距离法的相似度得分。例如“宇宙与空间探索”与“宇宙空间探索”，计算编辑距离时需要删掉第一个字符串中的“与”得到第二个字符串，因此编辑距离为 1，相似度得分为 6/7。

使用 Jaccard 算法计算时，可以选择使用单个汉字作为集合元素进行计算，也可以选择使用分词对文本进行切分，再将词作为集合进行计算。计算结果 J（A,B）即为 Jaccard 算法的相似度得分。还是以“宇宙与空间探索”与“宇宙空间探索”为例，若集合使用汉字作为元素，则相似度得分为 6/7，若使用分词进行切分，集合分别为【宇宙、与、空间、探索】和【宇宙、空间、探索】，相似度得分为 3/4。

使用 Word2Vec 算法计算时，先使用分词对两个文本进行切分，在将所有词转换为向量并求它们的平均值，计算两个文本向量的余弦相似度作为相似度得分。

相似度计算的代码如图 4-5 所示：

```
model_file = 'E:\\word2vec\\news_12g_baidubaike_20g_novel_90g_embedding_64.bin'
model = gensim.models.KeyedVectors.load_word2vec_format(model_file, binary=True)

def edit_distance(s1, s2):
    return 1 - distance.levenshtein(s1, s2)/len(s2)

def jaccard_similarity(s1, s2):
    cv = CountVectorizer(tokenizer=lambda s: list(s))  # 转化为TF矩阵
    corpus = [s1, s2]
    vectors = cv.fit_transform(corpus).toarray()
    numerator = np.sum(np.min(vectors, axis=0))  # 求交集
    denominator = np.sum(np.max(vectors, axis=0))  # 求并集
    return 1.0 * numerator / denominator  # 计算Jaccard系数

def vector_similarity(s1, s2):
    def sentence_vector(s):
        words = jieba.lcut(s, cut_all=True)
        add = 0
        v = np.zeros(64)
        for word in words:
            try:
                v += model[word]  # Word2Vec
            except:  # 异常处理
                try:
                    v += model[word[0]] + model[word[1]]
                    add += 1
                except:
                    add -= 1

        v /= len(words) + add
        return v

    v1, v2 = sentence_vector(s1), sentence_vector(s2)
    return np.dot(v1, v2) / (norm(v1) * norm(v2))  # 余弦距离
```

图 4-5 文本相似度计算方法的代码

Figure 4-5 Code for text similarity calculation method

4.2.5 缩小检测区间

在 4.2.2 节中，本文通过日志文件中的记录将每道题目的文本相似度检测区间从全部的录音文本缩小到了对应的题目中录音文本中，初步减轻了后续相似度检测的复杂度。经过对调查问卷项目的深入了解和录音文本的分析，本文发现了可以进一步缩小检测区间的方法。

上一节中提到了调查问卷的大题题目是由题干、题项与选项组成的，根据要求，我们需要将题目中这三类文本与分别于该题目对应的整段录音文本进行相似度计算。但我们发现，题目中的这三类文本在实际的调查过程中有潜在的顺序。调查员在进行调查时，通常需要先对题目的题干进行朗读，随后会按照题项的顺序，从 1 号题项开始依次对受访者进行提问，同时会在某一时刻朗读出题目的选项。因此顺序为：题干->题项 1->题项 2->…题项 n（选项待定）。

根据以上顺序，我们可以先计算题干与整段文本的相似度大小，并确定题干在整段文本中的位置，然后在计算题项 1 的相似度大小时，我们便可以从题干结束

的位置开始向后检测，同理题项 2 会从题项 1 的结束位置向后开始检测，以此类推。由于选项具有不确定性，因此建议从题干结束后检测。

为了防止极少数问卷没有按照顺序进行调查，我们会设置一个不同于审核所用阈值的较高阈值，若题项、或选项使用此方法时计算出的相似度低于这个阈值，则会对区间内未检测的文本重新进行检测。

这样做可以有效的缩小用于文本相似度计算时的区间，减少了不必要的计算，增加了系统审核时的效率。

4.3 图像分析模块具体设计

按照要求，调查过程需要在调查员和受访者一对一的环境中进行，并且实际中受访者的性别需要与调查所填的信息一致，因此本节给出 3.5 节中的图像模块的具体实现，基本原理是利用了人脸检测和性别二分类方法进行图像分析检测。

4.3.1 人脸数量检测

从调查问卷信息的压缩文件中可以解压出以“pic-X.jpg”命名的图片文件，这些文件是调查过程中进行随机拍摄的。考虑到调查需要在调查员和受访者一对一的环境中进行，而实际情况可能会出现在尚未准备好时进行拍摄，导致有些图片没有拍全等情况，本文算法将对所有拍摄图片进行人脸检测，检测时无须所有照片中都出现两张人脸，只要其中一张图片出现两张人脸即可，否则审核结果不通过。但由于要保证一对一的环境，因此倘若照片中检测出三张或以上的人脸，则会认定非一对一的环境，审核结果为不通过，并指出相应的图片并标注出其中的人脸位置。

本文使用开源的 MTCNN 卷积神经网络进行人脸检测，该网络可以输出识别到的人脸在图片中的位置列表，我们可以通过列表的长度判断图片中人脸的个数，以此来进行审核。

4.3.2 受访者性别检测

为了方便进行性别二分类判断，本文基于 CNN 卷积神经网络设计了性别二分类模型，并使用调查问卷的历史图片中截取出的人脸数据进行训练。

由于本文使用的服务器内存、算力有限，因此使用的 CNN 网络应在保证一定分类效果的情况下尽可能的简化模型。本文采用了比较经典的 3 卷积层+全连接层+sigmoid 层的性别二分类模型，可以保证该模型在本文服务器中正常运行。

模型的部分代码如图 4-6 所示：

```
x_input = tf.reshape(images_input,[-1,112,92,3])  # reshape图片

w1 = weight_init([3,3,3,16]) # 卷积核3*3*3 16个 第一层卷积
b1 = bias_init([16])
conv_1 = conv2d(x_input,w1)+b1
relu_1 = tf.nn.relu(conv_1,name='relu_1')
max_pool_1 = max_pool2x2(relu_1,'max_pool_1')  # Max pooling池化层

w2 = weight_init([3,3,16,32]) # 卷积核3*3*16  32个  第二层卷积
b2 = bias_init([32])
conv_2 = conv2d(max_pool_1,w2) + b2
relu_2 = tf.nn.relu(conv_2,name='relu_2')
max_pool_2 = max_pool2x2(relu_2,'max_pool_2')  # Max pooling池化层

w3 = weight_init([3,3,32,64]) # 卷积核3*3*32  64个  第三层卷积
b3 = bias_init([64])
conv_3 = conv2d(max_pool_2,w3)+b3
relu_3 = tf.nn.relu(conv_3,name='relu_3')
max_pool_3 = max_pool2x2(relu_3,'max_pool_3')  # Max pooling池化层
f_input = tf.reshape(max_pool_3,[-1,14*12*64])

f_w1= fch_init(14*12*64,512) # 全连接第一层
f_b1 = bias_init([512])
f_r1 = tf.matmul(f_input,f_w1) + f_b1
f_relu_r1 = tf.nn.relu(f_r1)  # 激活函数
f_dropout_r1 = tf.nn.dropout(f_relu_r1,drop_prob)

f_w2 = fch_init(512,128)  # 全连接第二层
f_b2 = bias_init([128])
f_r2 = tf.matmul(f_dropout_r1,f_w2) + f_b2
f_relu_r2 = tf.nn.relu(f_r2)  # 激活函数
f_dropout_r2 = tf.nn.dropout(f_relu_r2,drop_prob)

f_w3 = fch_init(128,2)  # 全连接第三层
f_b3 = bias_init([2])
f_r3 = tf.matmul(f_dropout_r2,f_w3) + f_b3

f_sigmoid = tf.nn.sigmoid(f_r3,name='f_sigmoid')  # 输出层
loss = tf.reduce_mean(tf.reduce_sum(abs(labels_input-f_sigmoid)))  # 损失函数
optimizer = tf.train.AdamOptimizer(learning_rate).minimize(loss)
```

图 4-6 性别二分类网络的代码

Figure 4-6 Code of the gender binary classification network

由图 4-6 可知，模型由输入层、卷积层、池化层、全连接层、激活函数层和输出层六部分组成：

第一部分为输入层，将调查问卷中的图片用 MTCNN 进行人脸检测，并且根据人脸的位置，将人脸的图像从图片中截取出来，并且将图片大小调整为 92*112 像素，作为模型的输入。

第二部分为卷积层，分别使用 16、32、64 个卷积核对图片进行卷积运算，核的长宽大小为 3*3，核的深度与上一层中的数据深度相同，目的是多方位的获取图片中的有用信息。

第三部分为 Max pooling 池化层，池化是降采样的一种形式，用于减小输入特

征的大小，减少神经网络运算时的参数和运算量，一定程度上可以防止过拟合的发生。

第四部分为全连接层，全连接层会将卷积层与池化层输入的数据做扁平化处理，由于扁平化后数据的维度较大，因此还会使用多个全连接层对数据进行降维。

第五部分为激活函数层，由于线性函数的表达能力不足，因此激活函数层会对全连接层的数据进行非线性处理，引入非线性的特征，提升神经网络的表达能力，本文使用了 relu 作为激活函数。

第六部分为输出层，分类问题的输出层一般用 softmax 函数或者 sigmoid 函数，由于模型属于性别二分类模型，因此使用 sigmoid 函数作为模型的输出层更为合适，模型根据输出结果的损失函数及优化器进行训练。

将模型训练好后，我们就可以对问卷中检测到两个人脸的图片进行性别检测，获得两个人的性别信息，并且通过与日志文件中的记录得到调查员与受访者两人的性别组成（如两男、两女、一男一女）作对比，查看图片中的性别组成与记录是否一致，以此来对问卷进行审核。

这里要注意的是，由于部分图片背景光线较暗、图片质量较差等缘故，可能存在部分图片中的人物性别识别不准的情况。因此本文采用了“多图验证”方式，即对同一问卷中的多张图片进行性别组成识别的结果一致时才能进行性别检测，这样做可以避免神经网络对个别图片性别识别错误时造成的误判，提升算法的准确率。

4.4 地理位置分析模块具体设计

本节给出 3.6 节中的地理分析模块的具体设计，其基本思想是利用调查问卷日志文件中包含的 GPS 位置信息，计算调查地点与目标居委会及附近的居委会的距离，并利用这些距离间的关系找出潜在的没有按照规定进行调查的不合格问卷，并列出错误信息，以便相关人员进行核实。

调查员使用 APP 进行调查时，APP 会使用 GPS 定位获得调查地点的地理位置信息，并将其记录在日志文件中。

APP 记录 GPS 信息时会每隔 10s 左右对位置进行一次检测，目的是防止 GPS 设备在定位时信号收到多种因素制约而造成的偏差。本文中对 GPS 信息的获取，会收集所有 GPS 的经纬度信息，并计算它们的平均值作为最终的调查地点的经纬度坐标。

记录的日志文件如图 4-7 所示：

```
2019-06-16 16-29-42,Latitude:38.051427980828834;   Longitude:114.50680083929119;&&
2019-06-16 16-29-52,Latitude:38.05179600149253;    Longitude:114.5070250171497;&&
2019-06-16 16-30-04,Latitude:38.05179375344004;    Longitude:114.507059105675;&&
2019-06-16 16-30-15,Latitude:38.051727470433605;   Longitude:114.50699705585035;&&
2019-06-16 16-30-25,Latitude:38.051727470433605;   Longitude:114.50699705585035;&&
2019-06-16 16-30-35,Latitude:38.05158229152886;    Longitude:114.50691410776341;&&
2019-06-16 16-30-47,Latitude:38.05161443615092;    Longitude:114.50690787750966;&&
2019-06-16 16-30-57,Latitude:38.05162453537977;    Longitude:114.50691846672298;&&
2019-06-16 16-31-07,Latitude:38.051547821714315;   Longitude:114.50697370432724;&&
2019-06-16 16-31-17,Latitude:38.05155996904733;    Longitude:114.50697154470703;&&
2019-06-16 16-31-27,Latitude:38.051615908541464;   Longitude:114.50694292450066;&&
```

图 4-7 日志中记录的 GPS 信息

Figure 4-7 GPS information recorded in the log

我们将通过对调查问卷编号的分析，得到调查的目标居委会。调查问卷编号由两部分组成，第一部分为调查目标居委会的编码，第二部分为调查进行的具体时间，其中，居委会编码的前三位代表省级编码，4-6 位为代表市级与区县编码，7-9 位为镇级/街道办编码，最后三位为居委会/村委会编码，编码会根据各省或直辖市的城市/农村结构而定，每三位划分一个级别。例如编号为“130105007016”的居委会代表“河北省石家庄市新华区中的一个对应的居委会”。

在对调查员作弊行为的分析中，我们发现有些调查员并没有按照要求前往目标居委会进行调查，而是选择了目标居委会附近范围内的其他居委会，这些居委会一般会处于同一个区/县中。因此本文将与目标居委会前六位编码相同的所有居委会作为调查员可能错误前往的附近居委会。

我们将使用百度地图 API 查询的方式事先获得所有调查地区中居委会的 GPS 地址，并将居委会编号和经纬度信息的键值对保存在“.pkl”文件中方便计算距离时进行调用。

在两地之间的距离时，会使用两地的经纬度坐标来进行计算，公式[32]如下：

$$d = 2 \cdot arcsin\sqrt{\sin^2\left(\frac{\delta_1 - \delta_2}{2}\right) + cos\delta_1 \cdot cos\delta_2 \cdot sin^2\left(\frac{\alpha_1 - \alpha_2}{2}\right)} \qquad (4-2)$$

其中，两点经纬度坐标分别为(α_1,δ_1)和(α_2,δ_2)为两点之间的距离。

我们首先查找出目标居委会的经纬度坐标，根据日志中记录的调查地点经纬度坐标，计算出调查地点与目标居委会的距离，记为 d1。再利用目标居委会的编号，找出与前六位编号相同的居委会，在键值对中查询出他们的经纬度坐标，计算出这些居委会与调查地点的距离，并取其中的最小值记为 d2。

最后将 d1 的值与 d2 的值作对比，由于通常情况下，调查地点应该距离目标居委会最近，而离其他居委会较远，因此 d1 应该小于 d2。本文利用这点，设计了一个可疑度得分，记为 d1-d2（单位 KM），如果 d1 小于 d2，则该得分小于 0，问

卷地理位置不可疑，如果 d1 大于 d2，则表示目标居委相较于附近的居委会离调查地点较远，d1-d2 的值越大，说明问卷地理位置越可疑，可疑就此设定阈值，超过阈值则认定该份问卷存在地理位置作弊的情况。

对于少部分因调查员个人或者设备问题造成 GPS 信息缺失的问卷，由于不能获得具体的位置信息，也会被系统判为不合格问卷，并将原因记录为“GPS 信息缺失”，报告给相关负责人员。

4.5 本章小结

本章的第一节主要给出了使用 API 接口实现了原有后台管理系统与本文审核算法的连接，第二节重点提出了使用静默检测、语音识别及本相似度计算对问卷进行审核的方法，第三节给出了通过 MTCNN 进行人脸检测来识别调查环境是否合格以及通过性别二分类 CNN 网络审核受访者性别信息是否正确，第四节使用日志中对 GPS 信息的记录以及各居委会的位置对调查员是否前往了正确的调查地点进行审核。

5 系统测试结果及分析

本章分为 7 个小节，第一节首先给出了实验的本地与服务器环境，第二节给出了对实验数据的统计，第三节给出了使用文本相似度进行审核算法的统计及结果，第四节展示了最终的系统展示界面，第五节给出了了图像、地理位置模块的结果信息，第六节对系统进行了性能测试，第七节为本章小结。

5.1 实验环境

本文设计系统的实验测试阶段在本地 windows 系统的计算机上完成，生产阶段部署在 Linux 系统的阿里云轻量级应用服务器中。具体环境如下：

(1)操作系统：windows 10（本地）； Ubuntu 16.04（服务器）

(2)处理器：2.6 GHz Intel Core i7-6700 HQ（本地）； 2.3 GHz Intel Xeon E5-2630（服务器）

(3)内存大小：8.00 GB（本地）； 4.00 GB（服务器）

(4)开发环境：Pycharm、Jupyter Notebook、sublime_text、HBuilder X

(5)开发语言：Python、PHP、HTML（JS+CSS）

(6)主要的工具包：Tensorflow、Gensim、Jieba、Pypinyin、Pandas、Yii2 等

5.2 实验数据统计

5.2.1 数据集

本文共收集了科普机构提供的 2019 年河北地区的 4341 份调查问卷，其中包含人工审核后判定为音频存在问题的不合格问卷共计 664 份，合格与不合格比例约为 5.5:1。

科普机构的“复审后废卷明细”提供了这 664 份问卷中每一道题目的题干、题项、选项是否合乎调查规范的标签，每份问卷共有 81 道小题的统计，共计 53784 道小题。本文针对录音文件的文本相似度检测算法就是以这 53784 道小题作为标签来衡量算法效果。

5.2.2 数据统计

(1)录音时长：在早期的人工审核中，审核人员通常会较为主观的将调查问卷的录音时长信息进行初审，通过录音长短来判断问卷是否合格。调查问卷中合格与不合格问卷的录音时长统计如图 5-1 所示：

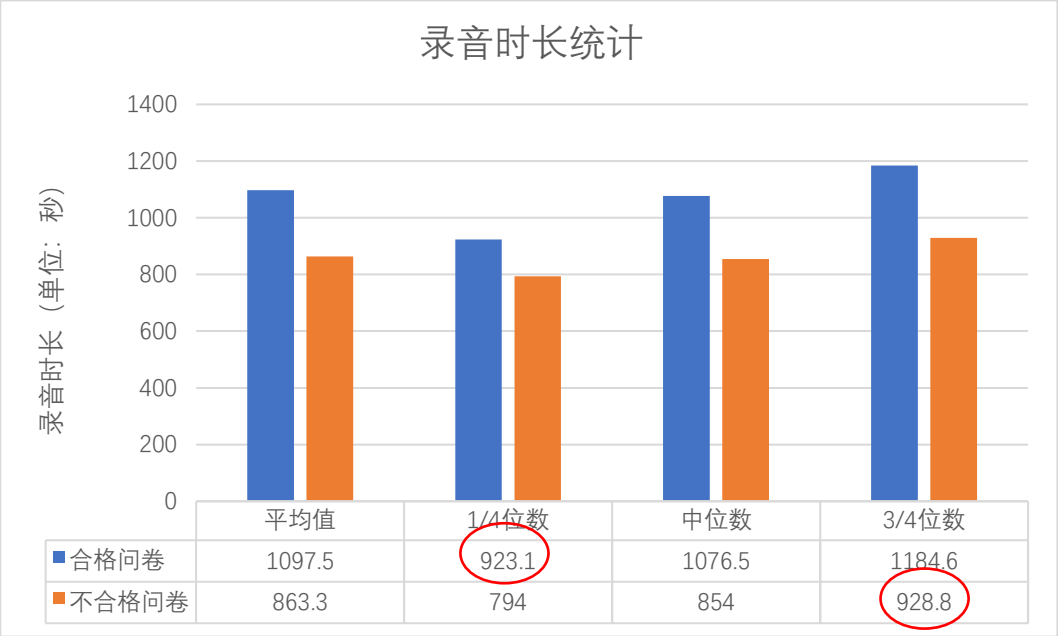


图 5-1 录音时长统计

Figure 5-1 Recording duration statistics

从图 5-1 中可以看出，合格问卷的录音时长的平均值、1/4、3/4 位数及中位数均高于不合格问卷，说明问卷的录音时长与问卷是否合格有较强的相关性。但同时我们注意到，不合格问卷的录音时长 3/4 位数略高于合格问卷的 1/4 位数，说明有部分不合格问卷的录音时长大于某些合格问卷，如果只凭借录音时长来进行问卷的审核，会对一些情况特殊的问卷造成一定的漏判或者误判。因此我们不能仅从录音时长来判断问卷是否合格，需要引入对语义的分析。

(2)静默检测统计

由于录音中会因调查员对 APP 的操作或者受访者回答问题时的思考而陷入短暂静默，这些静默片段不含有有用信息，在调用接口时会产生不必要的成本，因此本文系统会对去除录音的静默片段，保留非静默片段，问卷中录音的静默时长与非静默时长如图 5-2 所示：

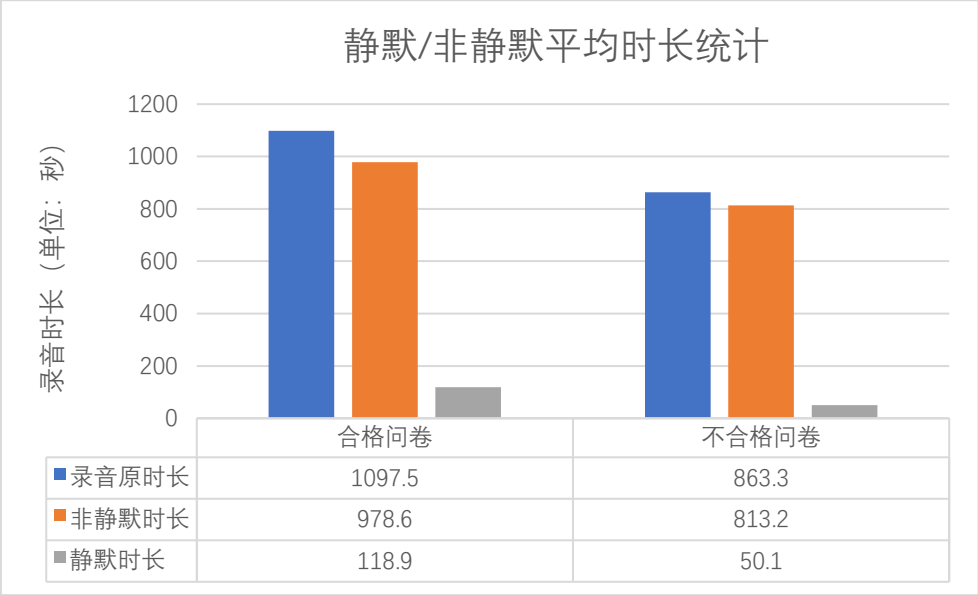


图 5-2 静默/非静默平均时长统计

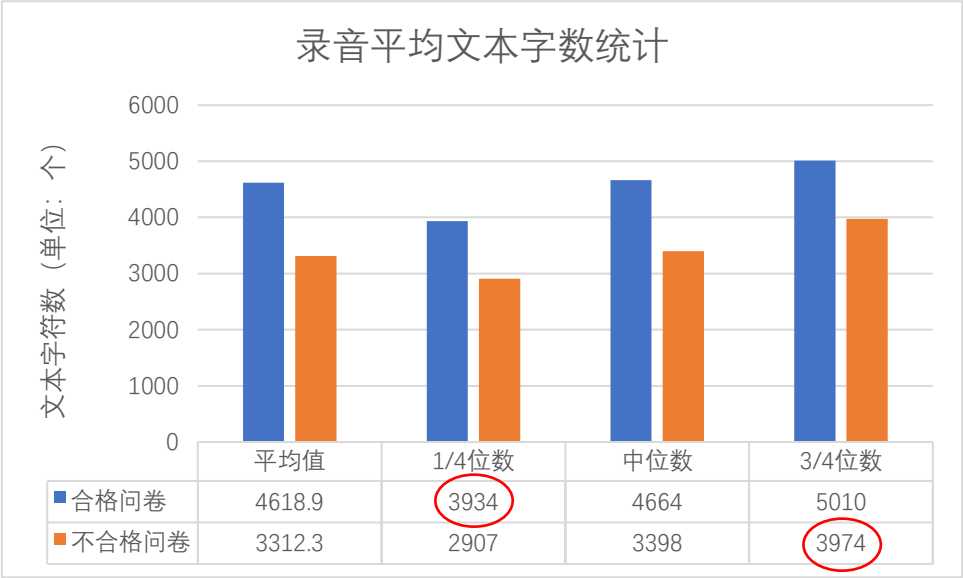
Figure 5-2 Quiet / non-quiet average duration statistics

从图中我们可以看出，合格问卷中检测出的静默时长占总时长的约 10%，不合格问卷中的静默时长占总时长的约 6%。综合合格问卷与不合格问卷的比例，在所有问卷中，静默时长约占总时长的 9.4%，这说明使用静默检测后，系统可以节省约 9.4%的接口调用成本。

值得注意的是，图中合格问卷与不合格问卷的原时长平均值相差约 20%，而合格问卷的静默时长却比不合格问卷的静默时长高了近 140%，这说明合格问卷相比于不合格问卷有着更大比例的静默时长。由于静默时长通常是留给受访者思考的时间，因此我们可以认为合格问卷中调查员给予了受访者充足的思考时间，相反不合格问卷的调查员可能为了节约时间成本，催促甚至诱导受访者尽快完成答题，严重影响了问卷收集到信息的质量。

(3)录音文本字数统计

通过语音识别接口，我们将录音的音频文件转成了对应的录音文本信息，录音文本字数统计如图所示：



5-3 录音平均文本字数统计

Figure 5-3 Recording average text word count

图中我们可以看出，合格问卷的文本字数的平均值、1/4、3/4 位数及中位数均大于不合格问卷。从不合格问卷的文本数的 3/4 分位数大于合格问卷的 1/4 分位数来看，说明部分不合格问卷中调查员与受访者进行了足够多的交流，但由于调查员并没有按照调查规范进行调查并朗读出相应的题目，因此也被审核为不合格。

与录音时长相似，录音文本字数的多少也与问卷是否合格有较强的相关性，但仅依靠文本字数对问卷进行审核仍然不妥，需要进行文本相似度分析对录音的语义进行审核，才能尽可能的避免误判和漏判。

我们将在下一节中介绍文本相似度匹配对录音进行语义审核的统计结果，以及通过阈值判断的审核结果与实际结果的比较。

5.3 文本相似度分析

5.3.1 文本相似度统计

通过对录音文本与目标文本的相似度计算，我们获得了问卷 81 道小题每道小题的相似度值，取这些值的平均值作为该份问卷最终的文本相似度值，图 5-4 中使用的相似度计算参数为：

文本相似度算法：Jaccard 算法；窗口大小：目标文本字符串长度+1；窗口移动步长：1；是否启用汉语拼音修正：是。

分别对合格问卷与不合格问卷进行了文本相似度统计，统计结果如下：

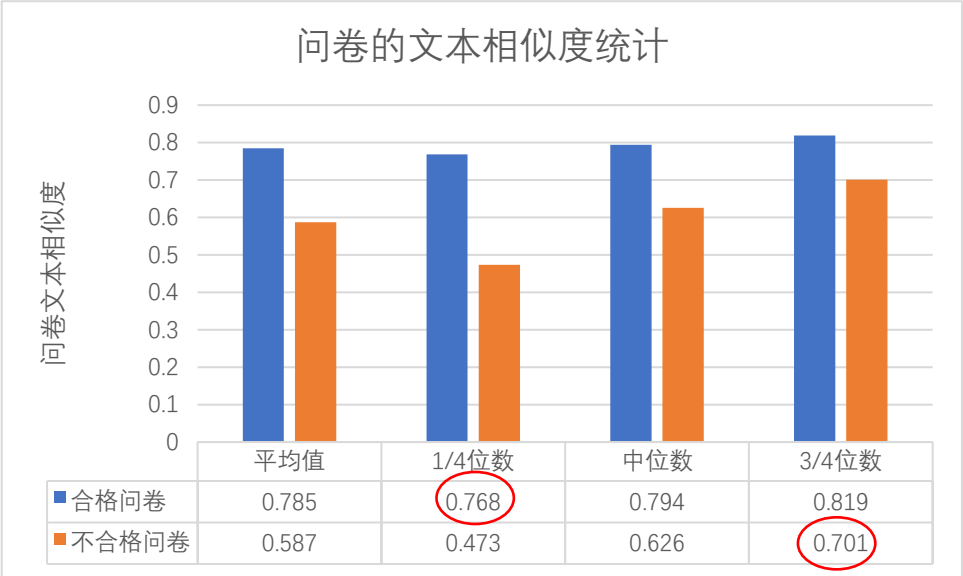


图 5-4 问卷的文本相似度统计

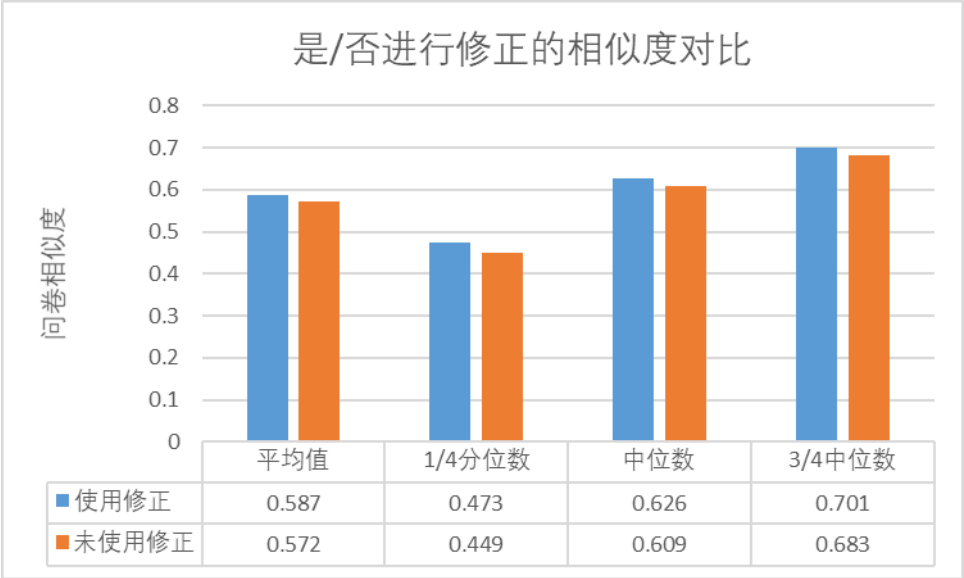
Figure 5-4 Questionnaire text similarity statistics

从图中可以看出，合格问卷的文本相似度平均值、1/4、3/4 位数及中位数均大于不合格问卷，且与问卷录音时长和文本字数不同的是，合格问卷的 1/4 分位数也不合格问卷的 3/4 分位数高了约 1/10。从数据中可以看出，绝大多数的合格问卷的文本相似度大于不合格问卷，两者间存在大致的边界。这说明充分利用了录音中语义信息的文本相似度分析算法的效果强于直接使用录音长度或字数进行分析的算法。

5.3.2 汉语拼音修正效果

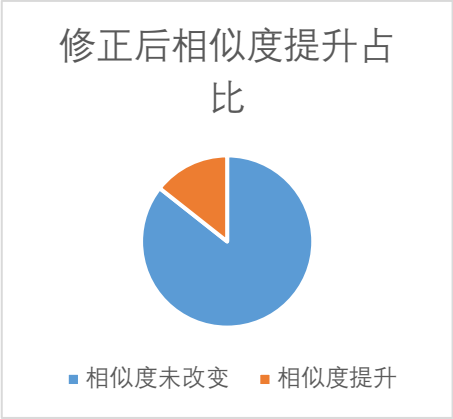
本文使用了一种基于汉语拼音的方法对录音转写后的文本进行了修正，通过这种做法可以使得修正后的文本内容更接近调查中实际的对话内容，从而提高计算出的文本相似度的值，使得一部分因为语音转写不准导致的误判被纠正过来，提高了算法的准确性。

以数据集中的不合格问卷为例，基于 Jaccard 相似度算法的修正效果如图 5-5 所示：



a) 修正前后数据对比

a) Comparison of data before and after correction



b) 相似度提升占比

b) Proportion of similarity increase

c) 审核结果改变占比

c) Changes in audit results

图 5-5 汉语拼音修正效果

Figure 5-5 Chinese pinyin correction effect

图 a 中显示了使用汉语拼音进行修正后，进行 Jaccard 算法计算所得的问卷相似度的平均值、1/4、3/4 位数及中位数均得到了小幅度提升，提升了 0.015 左右。图 b 中展示了修正后相似度提升的题目占所有题目的 14.3%，即汉语拼音修正法对 14.3%的题目起到了作用，平均每道题目的相似度提升了 0.1 左右。图 c 中显示了在这些相似度得到了提高的题目中，有 11%的题目（这些题目原本为合格题目但被误判为不合格）在提升过程中超过了问卷设定的阈值（0.4），使得原本审核为不合格的结果转变成了合格。

综合来看，使用汉语拼音进行修正后，使得全部问卷中 1.5%的题目的审核结

果从不合格变成了合格，即提升了 1.5%的正确率，减少了误判的发生。

5.3.3 算法评价指标及结果

本文对系统的算法评估使用了从科普机构获得的 664 份不合格调查中抽取的 500 份问卷，每份问卷针对其中出现 81 道小题有着明确的正确或错误的标注，因此评估主要涉及到对这 500×81 共计 40500 道小题的相似度打分。

本文以合格的小题为正例，不合格的小题为负例。其中 True positive(TP)代表“真正例”，将合格的小题预测为合格的个数；False positive(FP)代表“假正例”，将不合格的小题预测为合格的个数；False negative(FN)代表“假负例”，将合格的小题预测为不合格的个数；True negative(TN)代表“真负例”，将不合格的小题预测为不合格的个数。结合调查问卷的实际情况可以看出，TP、TN 值越高，FP、FN 值越低，对算法的评价越好。

本文对相似度算法的评估指标包括混淆矩阵、Precision、Recall、F1 Score 和 AUC 值，并针对 500 份不合格问卷中的 40500 道小题进行文本相似度计算，分别使用 Jaccard 相似度、编辑距离相似度和 Word2Vec+余弦距离相似度算法在使用汉语拼音文本修正的情况下进行计算，选取可以使 F1 Score 值尽可能高的阈值，分别取 0.40、0.35、0.83

算法表现如下表 5-1 所示：

表 5-1 算法评估结果

Table 5-1 Algorithm evaluation results

混淆矩阵	Jaccard	编辑距离	Word2Vec
TP	27025	27072	26083
TN	10467	10092	8783
FP	1025	1400	2709
FN	1983	1936	2925
算法指标			
Precision	0.963	0.951	0.905
Recall	0.932	0.933	0.899
F1 Score	0.947	0.942	0.903
AUC 值	0.956	0.951	0.920

从表中可以看出三种算法中最高的 Jaccard 算法，其 F1 Score 与 AUC 值高达 0.947 和 0.956，而苏论文中使用 Xgboost 模型得到的结果为 0.71 与 0.88，相较之下高了不少（尽管苏迪的论文使用的数据为 2018 年数据，而本文使用的是 2019 年

的数据，但是由于数据内容相近，比较还是有意义的）。这是因为苏的模型中只运用了录音时长的信息，没有涉及到深一层次的语义信息，因此依据 5.3.1 节中的分析，本文算法应该优于苏论文中的模型。

5.4 系统展示页面

本文设计系统为了方便科普机构相关人员对审核结果进行查看，使用了 HTML+Css+JavaScript 制作了一个有关录音文本相似度分析的页面。页面展示如图 5-6 所示：

图 5-6 系统展示页面

Figure 5-6 System display page

如图所示，页面上方位置记录了问卷中题目的编号，左侧浅绿色区域为 B 到 D 中大题的编号按钮，右侧黄色区域为大题对应的题干、题项及选项编号按钮，其中“D1-0”为选项编号。系统将对所有题目进行相似度计算，对于所有未通过阈值的题目，对应的编号按钮会变成红色，提示相关人员该题目存在错误。科普人员可以使用系统默认的阈值，也可以依据实际情况在后台对题目进行修改

下方位置记录了题目的相关信息，点击题目编号按钮后，左侧位置会出现当前题目所在区间的录音文本，并且将与当前题目相似度值最高的文本用红色标出。右侧会显示当前题目编号、相似度值、目标文本等信息，结合左侧的录音文本方便科普机构相关人员对审核结果进行查看。

5.5 图像及地理位置结果统计

5.5.1 图像部分

根据河北问卷中的 4341 份问卷中的随机拍摄图像进行人脸检测，取同一份问卷的 6 张图像中检测出人脸最多的数目作为该问卷人脸检测的结果，统计出的结果如图 5-7 所示：

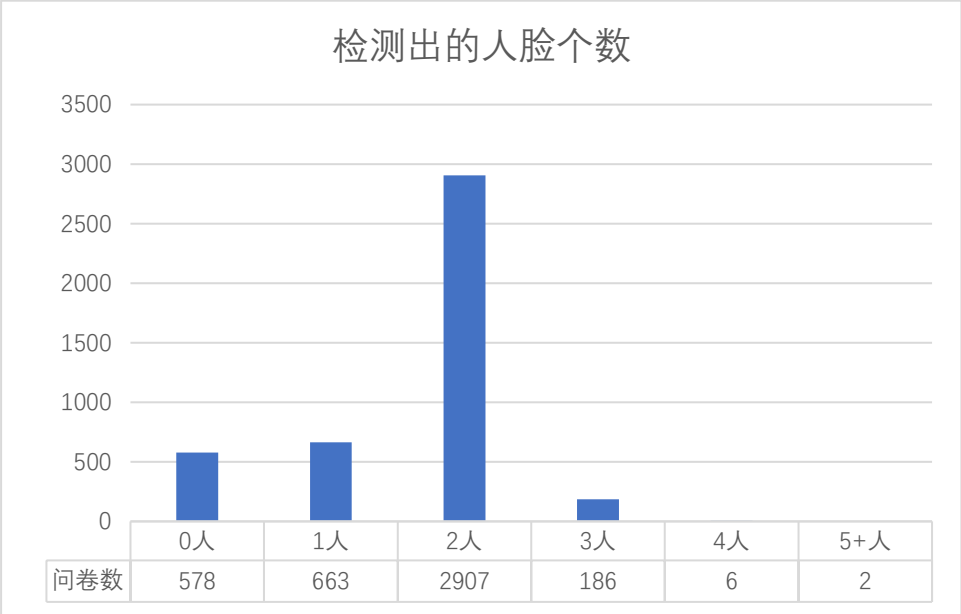


图 5-7 检测出的人脸个数

Figure 5-7 Number of detected faces

从图中可以看出，大部分问卷检测到了图像中最多出现了 2 个人，即符合调查员与受访者一对一的调查环境，还有一部分问卷识别出了 1 个人、3 个人及以上或没有识别到任何人，这些问卷则被认定为可疑问卷，会将错误报告给相关人员并在复审时重点处理。

对于检测到两个人的图像，系统会将人脸截取出来并使用卷积神经网络做性别二分类处理，并与实际调查员、受访者的性别组成加以验证。在 2019 年的 4341 份数据中，共有 2907 份问卷检测到了图像中出现调查员和受访者两个人，并使用“多图验证”方式对他们进行性别组合比对，最终共检测出了 39 份可能存在性别作弊的可疑问卷，后经人工核实，发现其中 34 份问卷情况属实，5 份问卷存在误判，准确率为 87%。而在无法使用“多图验证”方式进行检测的问卷中，经检验，仅有 2 份问卷出现了性别作弊问题。

由于目前现有的人工审核并没有在性别检测方面进行有效的审核措施，因此使用本文性别检测的方法可以发现约 0.8%存在性别作弊的问卷，提高了检测的准确率。

5.5.2 地理位置部分

计算调查地点与目标居委会距离记为 d1，找到目标居委会附近其他所有居委会，计算它们与调查地点距离的最小值记为 d2，计算 d1-d2 的值（单位 KM），其统计直方图与累计分布函数如图 5-8、5-9 所示：

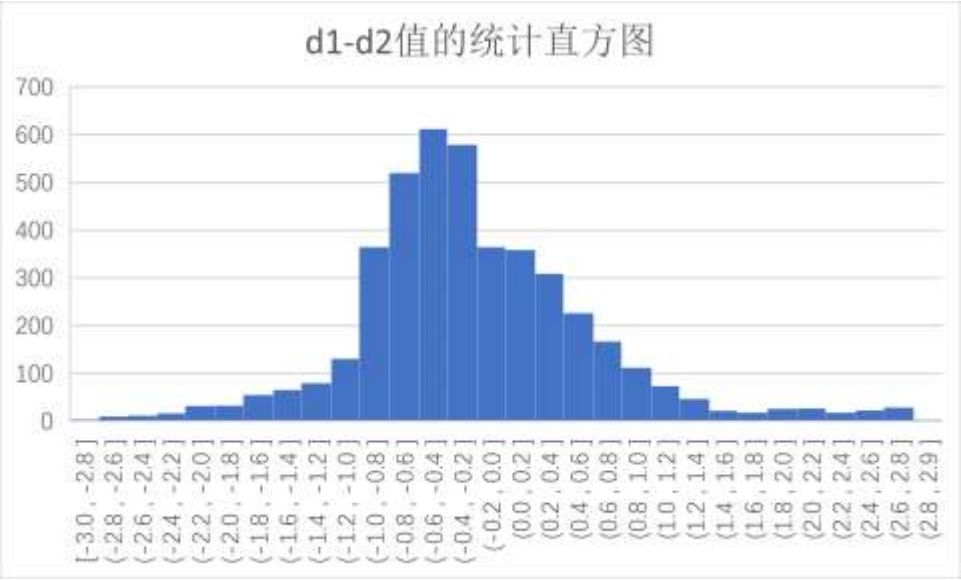


图 5-8 d1-d2 值的统计直方图

Figure 5-8 Statistical histogram of d1-d2 values

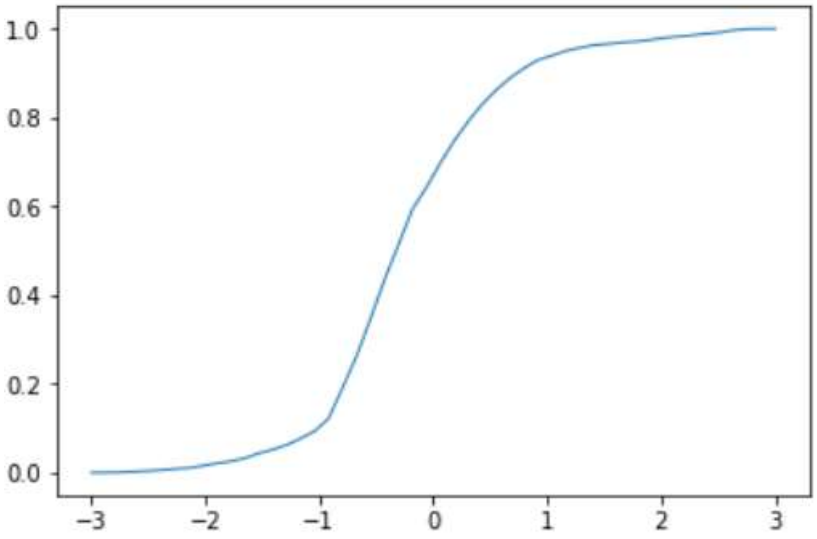


图 5-9 d1-d2 值的累积分布函数

Figure 5-9 CDF for d1-d2 values

从图中我们可以看出，对于大部分问卷来说，d1-d2 的值小于 0KM，即距离调查地点最近的居委会就是目标居委会。对于 d1-d2 的值大于 0KM 的问卷，存在比

目标居委会距离目标地点更近的其他居委会。d1-d2 的值越大，问卷可疑度越高。从累积分布函数（CDF）图来看，由于在 95%的问卷中，d1-d2 的值在 1KM 以下，因此本文将阈值设为 1KM，d1-d2 的值大于 1KM 的问卷将会被判为不合格问卷，并将可能错误前往的居委会编号及名称报告给相关人员以便后续进行复审。对于少部分因调查员个人或者设备问题造成 GPS 信息缺失的问卷，由于不能获得具体的位置信息，也会被系统判为不合格问卷。

在 2019 年河北省调查的 4341 份调查问卷中，除 5 份问卷由于调查员个人或者设备原因出现了 GPS 信息缺失的问题，剩余的 4336 份问卷中共有 78 份问卷被人工审核人员标注为“调查地点非抽样地点”，即出现地理位置不合格的问题，约占全部调查问卷的 1.7%，若使用本文地理位置部分的算法并以 1KM 作为阈值进行判断，可以全部识别出这 78 份不合格问卷，对于错误问卷的召回率为 100%。此外，由于现有的人工审核对于此类问题的审核力度较轻，只会标出一部分地理位置作弊较为明显的问卷，因而会有潜在的不合格问卷被审核为合格。使用本文的方法进行检测时，可以发现一些潜在的地理位置作弊的问卷，进一步加强问卷的审核力度。

5.6 系统性能测试

本文系统的生产环境使用到的算法模块主要部署在接口服务器上，因此我们需要统计各个模块的响应时间来计算系统的总响应时间，即接口处理一次请求所需要的时间。

我们可以通过事先将一部分算法所用的模型及数据加载到服务器内存中，这样避免每次调用接口时加载相同的资源影响运行速度。由于 Word2Vec 的模型较大，而服务器内存只有 4G，使用时会产生内存不足的情况，而且效果也较弱，因此 Word2Vec 算法将不被运用到最终的生产环境中。

其中，语音转写模块调用科大讯飞接口，不占用服务器负载，转写耗时依据科大讯飞服务器排队任务量浮动。

使用汉语拼音修正+Jaccard 算法效果最好，单次请求平均响应时间为 267.9 秒（合格问卷），210.6 秒（不合格问卷），其中语音转写模块占据了大部分时间，但由于该模块不占用系统自身服务器的负载，因此在接口并行处理多个请求时，可以快速的对多个问卷的该模块进行处理，系统的平均响应时间会大幅度缩短。

本文使用后台管理系统的自动请求脚本执行一次会向接口发送 10 份未审核的问卷请求。接口并行接收这些请求，从第一个请求发出开始计时，直到收到第十个响应，共花费 677 秒（10 份合格问卷）和 491 秒（10 份不合格问卷），可以在大约

10 分钟左右完成对请求的响应，平均每份问卷耗时 67.7 秒和 49.1 秒。

系统中各模块的响应时间如表 5-2 所示：

表 5-2 各模块响应时间

Table 5-2 Response time of each module

模块名称	相似度算法	是否启用汉语拼音修正	耗时（合格问卷）/秒	耗时（不合格问卷）/秒
去除静默			24.3	18.9
语音转写			201.5	166.2
区间对齐			<1	<1
相似度检测 1	编辑距离	是	18.0	9.6
相似度检测 2	编辑距离	否	5.6	3.6
相似度检测 3	Jaccard	是	38.1	21.5
相似度检测 4	Jaccard	否	26.1	16.6
图像模块			<2	<2
地理位置模块			<1	<1

5.7 本章小结

本章首先给出了本文系统的本地与服务器的实验环境，然后对本文数据集中的音频总长度、静默时长、文本字数进行了统计与分析，之后使用文本相似度计算方法对问卷的每个小题进行审核，统计了使用不同相似度计算方法的结果，并展示了供相关人员进行查看的审核结果展示界面，接着给出了图像及地理位置模块结果的统计数据，最后对系统各模块的性能及总体的运行效率进行测试。

6 结论

6.1 本文工作总结

调查问卷作为科普机构进行科普调查、掌握国民科学素质的重要工具，其收集到数据的可靠性在整个中国公民科学素质测评体系中起到了举足轻重的作用。然而对于调查问卷可靠性进行人工审核，需要花费大量的时间和精力，代价巨大。本文通过对问卷不合格原因及调查员的作弊行为进行分析，使用语音识别、文本相似度检测、人脸检测等技术，设计了一套调查问卷的自动化审核系统，提高了问卷的审核效率，并能提供详细错误信息协助问卷的复审工作，具体工作如下：

(1)通过语音识别技术，将调查过程中的录音信息转成文字信息，并根据相关题目的文本进行文本相似度分析，根据设定的阈值确定错误题目编号及类型，并使用静默检测、汉语拼音修正等方法提高系统的效率及准确率。

(2)针对问卷中随机拍摄的图像，使用人脸检测方法得出图像中的人数，并判断调查环境是否合乎要求。对检测出的人脸进行性别识别，判断调查员是否找到了对应性别的受访者进行调查。使用 GPS 功能对调查地点的地理位置进行分析，判断调查员是否按照要求前往了指定的居委会调查。

(3)为了保证系统可以实际运行，在阿里云服务器上部署了 Flask 接口连接算法，对原有的后台管理系统进行了修改，并提供了相关审核页面方便相关人员对错误原因进行查询。使用多线程并行处理、IP 地址检测等方法，保证了系统的有效性和可靠性。

6.2 未来工作展望

本文进行的研究工作虽然取得了一定的成果，但仍然有一些不足的地方，未来可以逐步的完善和改进，具体方向如下：

(1)由于文本相似度检测是针对目标题目文本没有被朗读出来这一错误情况进行的检测，而现实中存在少数调查员误导受访者进行答题的错误情况，这种情况在本文设计中并没有检测出来。因此可以根据经验总结出调查员误导受访者时使用的语句，对这些语句进行相似度分析或更深层的语义分析。

(2) 本文中使用汉语拼音进行文本修正的方法只涉及到两个字组成的词语，且需要有 3 个以上的声/韵母相同才能修正，条件相对苛刻。可以针对具体的题目文

本对该修正方法进行改进，使其可以准确的修正更多的文本，进一步提高可靠性。

(3)后台系统向接口发送请求后，发现接口所属服务器 CPU 使用率并非一直处于满载状态，因此会有一定的资源浪费，推测是由于调用语音转写接口的很长一段时间内并不需要使用本服务器进行计算所导致的。后期可以尝试通过修改请求的发送间隔，充分利用服务器的 CPU 算力，进一步缩短接口平均响应时间，提高系统的运行效率。

参考文献

[1] 汤书昆,王孝炯,徐晓飞.中国公民科学素质测评指标体系研究[J].科学学研究,2008(01):78-84.

[2] 王宇良, 戚敏. 科普调查问卷及其设计技巧的探析[J]. 科普研究, 2010, 1: 37-42.

[3] 王海坤, 潘嘉, 刘聪. 语音识别技术的研究进展与展望[J]. 电信科学, 2018, 34(2): 1-11.

[4] Murali M T S V, Murali V, Reddy B P. Human Face Detection & Segme Binary Patterns i[J]. 2019.

[5] Antipov G, Baccouche M, Berrani S A. Effective training of convolutional neural networks for face-based gender and age prediction[J]. Pattern Recognition, 2017, 72: 15-26.

[6] 苏迪. 基于机器学习的问卷可信度审核系统[D]. 北京交通大学, 2019.

[7] Bansal P, Kant A, Kumar S, et al. Improved hybrid model of HMM/GMM for speech recognition[J]. 2008.

[8] Hinton G E, Osindero S, Teh Y W. A fast learning algorithm for deep belief nets[J]. Neural computation, 2006, 18(7): 1527-1554.

[9] Larochelle H, Bengio Y, Louradour J, et al. Exploring strategies for training deep neural networks[J]. Journal of machine learning research, 2009, 10(Jan): 1-40.

[10]Hinton G, Deng L, Yu D, et al. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups[J]. IEEE Signal processing magazine, 2012, 29(6): 82-97.

[11]Graves A, Mohamed A, Hinton G. Speech recognition with deep recurrent neural networks[C]//2013 IEEE international conference on acoustics, speech and signal processing. IEEE, 2013: 6645-6649.

[12]Graves A, Jaitly N. Towards end-to-end speech recognition with recurrent neural networks[C]//International conference on machine learning. 2014: 1764-1772.

[13]讯飞开放平台 https://www.xfyun.cn

[14]Zhang S, Liu C, Jiang H, et al. Feedforward sequential memory networks: A new structure to learn long-term dependency[J]. arXiv preprint arXiv:1512.08301, 2015.

[15]常建秋, 沈炜. 基于字符串匹配的中文分词算法的研究[J]. 工业控制计算机, 2016,

29(2): 115-116.

[16]翟凤文, 赫枫龄, 左万利. 字典与统计相结合的中文分词方法[D]. , 2006.

[17]金宸, 李维华, 姬晨, 等. 基于双向 LSTM 神经网络模型的中文分词[J]. 中文信息学报, 2018, 32(2): 29-37.

[18]Xu J, Sun X. Dependency-based gated recursive neural network for chinese word segmentation[C]//Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). 2016: 567-572.

[19]结巴分词工具开源社区. https://github.com/fxsjy/jieba.

[20]哈工大语言云技术平台. http://www.ltp-cloud.com.

[21]陈二静, 姜恩波. 文本相似度计算方法研究综述[J]. 数据分析与知识发现, 2017, 1(6): 1-11.

[22]Yih W T, Meek C. Improving similarity measures for short segments of text[C]//AAAI. 2007, 7(7): 1489-1494.

[23]Masek W J, Paterson M S. A faster algorithm computing string edit distances[J]. Journal of Computer and System sciences, 1980, 20(1): 18-31.

[24]李晓, 解辉, 李立杰. 基于 Word2vec 的句子语义相似度计算研究[J]. 计算机科学, 2017, 44(9): 256-260.

[25]Mikolov T, Chen K, Corrado G, et al. Efficient estimation of word representations in vector space[J]. arXiv preprint arXiv:1301.3781, 2013.

[26]Jones P, Viola P, Jones M. Rapid object detection using a boosted cascade of simple features[C]//University of Rochester. Charles Rich. 2001.

[27]Viola P, Jones M. Robust real-time object detection[J]. International journal of computer vision, 2001, 4(34-47): 4.

[28]Jiang H, Learned-Miller E. Face detection with the faster R-CNN[C]//2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017). IEEE, 2017: 650-657.

[29]Ma M, Wang J. Multi-View Face Detection and Landmark Localization Based on MTCNN[C]//2018 Chinese Automation Congress (CAC). IEEE, 2018: 4200-4205.

[30]Zhang K, Zhang Z, Li Z, et al. Joint face detection and alignment using multitask cascaded convolutional networks[J]. IEEE Signal Processing Letters, 2016, 23(10): 1499-1503.

[31]Ma J, Chen L, Gao Z. Hardware implementation and optimization of tiny-yolo network[C]//International Forum on Digital TV and Wireless Multimedia Communications. Springer, Singapore, 2017: 224-234.

[32]Fan Dongwei, He Boliang, Li Changhua, Han Jun, Xu Yunfei, Cui Chenzhou. Research on Spherical Distance Computation and Accuracy Comparison. Astronomical Research and Technology, 2019, 16(1): 69-76.

作者简历及攻读硕士/博士学位期间取得的研究成果

一、作者简历

刘翰文，男，1996 年 8 月生。2014 年 9 月至 2018 年 7 月就读于北京交通大学电子信息工程学院通信工程专业，取得工学学士学位。2018 年 9 月至 2020 年 6 月就读于北京交通大学电子信息工程学院电子与通行工程专业，获得工学专业硕士学位。

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和取得的研究成果，除了文中特别加以标注和致谢之处外，论文中不包含其他人已经发表或撰写过的研究成果，也不包含为获得北京交通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并表示了谢意。

学位论文作者签名：　　签字日期：　　年　　月　　日

学位论文数据集

表 1.1： 数据集页

<table>
<tr><td>关键词*</td><td>密级*</td><td>中图分类号</td><td>UDC</td><td>论文资助</td></tr>
<tr><td>科普调查问卷</td><td>公开</td><td></td><td></td><td></td></tr>
<tr><td colspan="2">学位授予单位名称*</td><td>学位授予单位代码*</td><td>学位类别*</td><td>学位级别*</td></tr>
<tr><td colspan="2">北京交通大学</td><td>10004</td><td>工学</td><td>硕士</td></tr>
<tr><td colspan="2">论文题名*</td><td colspan="2">并列题名</td><td>论文语种*</td></tr>
<tr><td colspan="2">调查问卷自动化审核系统的实现及其优化</td><td colspan="2"></td><td>中文</td></tr>
<tr><td>作者姓名*</td><td colspan="2">刘翰文</td><td>学号*</td><td>18125030</td></tr>
<tr><td colspan="2">培养单位名称*</td><td>培养单位代码*</td><td>培养单位地址</td><td>邮编</td></tr>
<tr><td colspan="2">北京交通大学</td><td>10004</td><td>北京市海淀区西直门外上园村 3 号</td><td>100044</td></tr>
<tr><td colspan="2">工程领域*</td><td>研究方向*</td><td>学制*</td><td>学位授予年*</td></tr>
<tr><td colspan="2">电子与通信工程</td><td>信息网络</td><td>2 年</td><td>2020</td></tr>
<tr><td>论文提交日期*</td><td colspan="4"></td></tr>
<tr><td>导师姓名*</td><td colspan="2">赵永祥</td><td>职称*</td><td>副教授</td></tr>
<tr><td>评阅人</td><td colspan="2">答辩委员会主席*</td><td colspan="2">答辩委员会成员</td></tr>
<tr><td></td><td colspan="2"></td><td colspan="2"></td></tr>
<tr><td colspan="5">电子版论文提交格式　文本（ ）　图像（ ）　视频（ ）　音频（ ）　多媒体（ ）　其他（ ）
推荐格式：application/msword；application/pdf</td></tr>
<tr><td colspan="2">电子版论文出版（发布）者</td><td colspan="2">电子版论文出版（发布）地</td><td>权限声明</td></tr>
<tr><td colspan="2"></td><td colspan="2"></td><td></td></tr>
<tr><td>论文总页数*</td><td colspan="4">59</td></tr>
<tr><td colspan="5">共 33 项，其中带*为必填数据，为 21 项。</td></tr>
</table>