2017 年本科生科技创新创业训练项目

申 请 书

项目名称　基于大数据的智能问卷系统

项目类型　■创新训练　□创业训练　□创业实践

项目周期　■ 1 年　□ 2 年

项目负责人　赵吉彤

学　　号　41524122

项目立项学院/单位　计算机与通信工程学院

联系电话　13220101996

北京科技大学教务处

2016 年 10 月

填 写 须 知

一、项目分类说明

1．创新训练项目是本科生个人或团体，在师指导下，自主完成创新性实验方法的设计、设备和材料的准备、实验的实施、数据处理与分析、总结报告和科研论文的撰写等工作。

2．创业训练项目是本科生团队，在导师指导下，团队中每个学生在项目实施过程中扮演一个或多个具体的角色，通过编制商业计划书、开展可行性研究、模拟企业运行、进行一定程度的验证试验，撰写创业报告等工作。

3. 创业实践项目是在学校导师和企业导师共同指导下，采用前期创新训练项目成果或团队成员已有的研究成果，提出一项具有市场前景的创新性产品或者服务，以此为基础开展实际的创业实践活动。

二、申请书请按顺序逐项填写，填写内容必须实习求是，表达明确严谨。空缺项要填“无”。

三、创业类项目填写内容说明：项目简介主要说明项目来源、产品和服务介绍、项目优势、市场需求等；申请理由包括项目背景和意义、团队优势等；项目方案主要阐述执行思路；项目特色与创新主要说明项目特色和可行性；项目预期成果包括商业计划书、发表论文，调研报告，竞赛获奖、创建和运营公司等相关成果。

四、申请大学生创新创业训练计划项目团队的人数含负责人在内不得超过5人。

五、填写时可以改变字体大小等，但要确保表格的样式没有改变；填写完后用A4纸张双面打印，不得随意涂改。

六、申报过程有不明事宜，请向教务处教学科办公室联系咨询，电话62332202。

项目名称	基于大数据的智能问卷系统					
项目负责人	姓　名	性别	学　号	班　级	所在学院	联系电话
	赵吉彤	男	41524122	计 1501	计通	13220101996
项目组成员	夏超	男	41455111	计 1404	计通	17801048128
	孙天宇	男	41524109	计 1501	计通	18810580160
	无					
	无					

指导教师	姓　名	职　称	所在学院	联系电话
	王睿	副教授、副研究员	计通	13910148280
	无			

一、项目简介（500 字以内）

基于大数据的智能问卷系统是一项与各个相关领域直接合作的科研项目，传统的问卷系统一般先由相关人员设计问卷，然后印发纸质版并由相关人员派发和回收或第三方电子版问卷，这种形式的交互存在很多问题和隐患，且在数据分析方面尚有很大欠缺，因此基于大数据的智能问卷系统应运而生：

1.首先，我们通过优化纸质问卷，为目标人群开发出界面友好的问卷系统。我们会运用已有的数据，进行充分的机器学习和数据挖掘，提升目标人群在填写问卷时的用户体验；

2.其次，我们在后端服务的优化、数据模型和数据库的建立等方面做出了充足的考虑，为每一个目标人群建立足够庞大的亲朋关系网，努力将人性化的用户体验做到极致。

3.而且我们为问卷系统开发非常完备的管理员后台，除常规的数据库删改，还会有多维度、可拆分、深度定制的数据分析及展示模块，通过我们对已有并不断补充的数据分析，建立科学有效的数据分析模型，帮助调查人员对目标人群整体情况的把握，提供完备多角度的数据展示。

我们希望通过对各行各业对智能分析需求的整体把握和感知，把智能问卷系统做的深度可定制化、模型高复用性和数据多维度性。

在实际生产中，将多重问题模型算法高内聚低耦合，以人工最少的方式为调查方提供科学有效的分析数据。

二、选题背景(2000 字以内)

生活中很多行业的调查问卷，一般具有专业性、数据源高度保密性、权威性和科学性，而现有模式一般使用常规调查手段和分析手段，并不能很好的契合新兴互联网模式下的高效特点。所以我们与有关行业（例如目前已有合作的医疗行业）密切协作，采取机器学习和数据挖掘的模式，对现有数据进行分析处理，构建和完善数据处理模型。 现有问卷存在的问题有：

1. 问卷方式陈旧： 据我们了解，调研方因为种种原因，在某些调查研究类活动的实施上仍采取以前的方法，比如调查问卷人力发放和回收。 这对于前期分发问卷和后期收集整理数据是极为不利的，尤其是对于受调查目标人群的筛分、联系、回访，以及对数据的建模、分析、处理。

2. 难以深度定制： 因为问卷是由相关专家经研讨制定，所以存在普适性和复用性。但纸质问卷在更改和定制方面无法做到与移动问卷相媲美，而且其浪费纸质资源， 在运输过程中也容易损坏和丢失，在种种方面都有着巨大的市场空间待弥补。

3. 数据分析处理不够智能：现阶段调查问卷的处理方式是通过人工录入计算机或数据库，通过导出 Excel 表格进行进一步数据加工处理分析。 这种方式存在很大的人为主观因素以及不确定性，在多种数据模式下也难以做到高复用性。 或是通过第三方受到限制浏览部分数据，且数据展示方式比较原始。 同时，这种延迟性交互的效果比较传统，目标用户无法在第一时间得到内心预期的数据分析结果。

4. 集成化可复用性有待提高： 由于没有统一的信息源管理系统，不同医院、不同科室、不同调查人员对信息的录入检出的方式会存在很大差异， 这些差异导致了信息源的集成化程度低，分析方法可复用性低，甚至不可复用。

通过我们对这些需求的了解和调研，拟定了“基于大数据的智能问卷系统”的项目计划，我们本着科学性、权威性和高度拟合用户的初衷，为此项目制定了项目方案及进度安排。

三、项目方案（1500 字以内）

基于大数据的智能问卷系统主要着重以实现以下几方面功能：

1. 通过友好的界面从目标人群处采集尽可能详实可靠的信息，依托已有数据及后台统计分析分别对目标人群予以易于理解的反馈，对调查人员给予具有专业指导意义的反馈以及通过对数据集进行系统的模式识别得到具有实践意义的学术结论。

a) 首先，针对从目标人群处采集信息的过程，我们将对纸质问卷的内容及形式进行充分的优化，基于 Angular JS 及 Node JS 等技术以及精细的 UI 设计，为用户设计一款友好的交互界面，在提升用户体验的同时，更为我们从用户处采集到的信息的有效性提供了有效的保障。

b) 在后台，我们将依托现有数据训练科学有效的模型，以对新采集到的数据进行高效的分类并并入现有数据集。通过以现有数据集为基础的模式识别来为数据分析提供支持。我们将在现有数据对象以及新采集到的数据对象之间建立贝叶斯网络，并对各数据类进行相关性分析，然后结合决策树等模型对目标人群予以反馈。

2. 此外，考虑到人的心理状态及身体状态是不稳定的，所以我们需要对新采集到的数据的有效性进行进一步的保证。我们将通过 Web API 采集用户的设备的硬件传感器数据，主要对摄像头采集到的图像信息和麦克风采集到的语音信息进行分析。对于图像信息，我们将首先与现有成熟 API 相结合，首先得到一些基础数据。如，在多次问卷这个持续的过程中，我们可以通过第一次问卷时采集到的目标人群容貌信息来 进行身份识别，在填写问卷时通过将摄像头采集到的图像信息交由成熟 API 进行处理来确认用户身份，确保问卷由目标人群本人填写以保证真实性。

3. 在结合成熟 API 之外，我们还将依托现有信息及数据对图像信息进行进一步处理。如，我们可以通过非首次用户之前的信息来大致判断用户的身体状况的大致属性，然后将新采集到的图像信息与已知信息进行匹配，从而建立一个目标人群身体状况与面部图像（如脸色等）的关系集，再基于这个关系集进行横向比对从而实现对新采集到的首次用户的图像信息的贝叶斯分类，从而将新采集到的图像信息加入到现有 贝叶斯网络中。

三、项目方案（1500 字以内）

4. 对于语音信息，为了增加我们能够获得的信息，我们将推荐用户进行语音输入。一方面提升了用户体验，使用户填写问卷的过程更加方便快捷，另一方面也间接地有利于用户所填写信息的真实有效。对于采集到的语音信息，我们同样将首先依托现有成熟 API 进行初步处理，将语音信号转换成文字信息。另一方面，我们也将对所得语音信息进行进一步处理。

5. 我们将采取与图像处理类似的思路，依托现有数据以及已知信息在数据集间建立关系网络，通过对语音信号的进一步处理来分析目标人群的情绪及身体状况，并结合对其他因素的相关性分析来评估目标人群所输入信息的有效性。如，我们可以借助非首次填写问卷的目标人群的既有数据来分析目标人群新输入信息的有效性，如身高等因素的可变性很小，若目标人群输入信息变化较大，则我们需要关注信息是否真实有效。

6. 对于问卷填写者，我们将建立相关数据模型以确保其填写真实性和是否认真。如在问卷中加入一个此时天气的问题，将采集到的数据与权威数据进行比对，从而对目标人群填写问卷的认真程度进行评估。

7. 在对所有可获得数据集进行综合系统的模式识别后，我们将分析每个具体对象的身体状况的主导影响因素，并结合医院方面调查人员建议给出一定的生活建议，以帮助目标人群更好地保持良好的身体状况。

8. 此外，我们还把将对所有数据进行分析后得到的主要结果进行数据可视化，并为调查人员方面提供一个友好的 UI 界面，使调查人员能够对基于大量数据分析得到的结果有一个比较深入的了解，以便更好的把握目标人群的病情。

9. 在保证实用性之外，我们还将与有关行业，例如医疗、教育等行业，进行密切合作，结合调查人员的专业知识对所有可获得数据集进行综合系统的非监督式学习，以期得到具有实践意义的学术成果。

四、项目特色与创新（500 字以内）

基于大数据的智能问卷系统的特色创新主要集中于信息采集过程以及通过后台数据分析实现的与用户的交互。

1. 首先，在数据采集过程中，我们将不仅仅依赖于用户的手动填写，而是通过浏览器 Web API 调取用户终端多维可用数据，并支持语音输入。对于从硬件传感器处获得的语音、图像等数据。

a) 一方面通过现有成熟 API 进行处理获得基础信息，如通过图像识别确认填写者是否为目标人群本人，通过语音识别将目标人群的语音输入转换为文字等

b) 另一方面，我们将对硬件传感器的数据进行进一步分析，如通过语音信号的语音语调判断目标人群的情绪，通过图像信号对目标人群的表情、脸色等分析目标人群的身体状况等。

2. 依托现有数据集，我们将通过概率统计分析以及建立相关因素贝叶斯网络等技术对目标人群所填信息的真实性进行评估。

a) 对于问卷填写者，我们将建立相关数据模型以确保其填写真实性和是否认真。如在问卷中加入一个此时天气的问题，将采集到的数据与权威数据进行比对，从而对目标人群填写问卷的认真程度进行评估。

3. 我们将利用现有的数据集选择合适的模型进行模式识别，通过训练出的模型来分析新的信息，在目标人群填写完问卷时予以反馈，如评估目标人群身体状况发展趋势，分析当前对目标人群身体状况影响较大的因素，对目标人群的生活方式给出相应建议等。

五、项目进度与安排（1000 字以内）

进度：

1．本项目自 2016 年 11 月开始，与 2017 年 1 月前完成第一版封版，其中第一版的预期成果和规模如下：

a) 调研有关行业需求，商讨数据关系
b) 搭建数据库，敲定数据模型
c) 完成调查问卷前端和后台服务的编写
d) 用户使用无明显障碍

2．2017 年 1 月到 2017 年 5 月，完成项目第二版封版：

a) 完成管理员后台管理的页面及功能
b) 依赖院方提供的可用数据，逐步开展机器学习和数据挖掘
c) 优化用户展示界面，将机器学习成果运用到展示页面
d) 期间完成 SRTP 项目的中期检查，做好相应单元测试和阶段性测试

3．2017 年 5 月到 2017 年，项目最终版封版：

a) 前端方面可能会用 ES6 和 VueJS 进行项目重构，力图做到“高内聚低耦合”，尽可能减小后期维护的复杂度
b) 完善管理员后台功能，及数据分析处理之后的展示页面
c) 利用机器学习结果，为不同用户深度定制科学的问卷系统
d) 完善相关医疗问卷系统的数据模型，建立多维度、深层次的模型，供后期二期开发
e) 将数据处理和机器学习得出的结论整理汇总，以论文形式发表

安排：

4．前期准备工作：

a) 人员方面：项目分工明确，前端和 UI/UX 方面赵吉彤负责，后端由夏超负责，机器学习及数据处理孙天宇主要负责。
b) 协商产品：项目实施过程，每周固定时间开展讨论会，并在会上约定相关开发规范，前后端接口等。该会频次依照项目进度弹性调整。
c) 开发合作：项目整体依托 git 进行开发过程中的团队合作，目前前端项目开源并托管在 GitHub(https://github.com/jeasonstudio/MammaryCancerPaper)，后端及机器学习部分闭源。

5．开发过程：

a) 及时沟通需求，规划开发周期，避免项目延期
b) 按时总结开发过程中遇到的困难疑惑和感悟，组内成员讨论并时时与导师交流沟通，认真学习专业知识
c) 在产品的安全性方面多加考虑，规避 xss、xrsf、sql 注入等前后端渗透漏洞

六、项目经费使用计划（500 字以内，作为评审参考，具体划拨经费方案见立项启动通知。）

该项目申请的经费主要包括下面几个用途：

一、 项目相关资料购买：其中包括购买学习前端、后端、数据挖据方面的书籍或视频教程的费用。此项预计 1000 元。

二、 沟通与调研：由于此项目涉及到部分信息来源的沟通确认，与相关院方的协调和需求定制，与各个相关科室的深度定制，所以此项预计费用 500 元。

三、 服务器、域名及其他服务：由于此项目在部署方面、开发进行中，需要有较高的配置服务器和其他如 CDN、域名解析等服务的支持。故预计此项费用为 2000 元/年。

四、 微信服务号开通费用。预计约 500 元。

五、 其他元器件的购买：此项预计费用 500 元。

一年周期项目合计费用约 4500 元。

七、项目成果形式

一、 完整的基于大数据的智能问卷系统：其包括结合数据处理后的前端用户界面、描述详尽的后台管理及数据展示界面、完备的后台服务及数据库数据模型。

二、 对于用户终端硬件的创新利用：会调用如摄像头、语音等硬件设备数据进行深度处理，丰富目标对象的数据模型。

三、 多维度的机器学习模型建构：通过项目过程中对大量已有数据的机器学习、数据分析，建立多维度、深度可定制化的数据模型，方便复用及二次开发。

四、 相关论文发表：项目研发后端及数据挖掘方面的成果讲以论文形式发表，并努力达到一定科研影响。

成果特征： 硬件□ 软件■ 研究报告■ 高水平论文■ 经济效益■

指导教师意见：

相比传统的调查问卷，本项目拟开发的利用移动端开发的智能问卷系统可以有效解决问卷不真实、采集不方便，分析不精准的问题，具有较好的创新性和实用价值，同时该项目已在前期完成 demo 开发，可行性较好，开发的系统或关键技术将有可能在医疗人群队列调查问卷中进行试用，故建议优先推荐该项目参加本科生科技创新创业训练项目。

签　名：　　年　　月　　日

学院意见：

是否同意立项：　　是　口　　否　口

推荐项目级别：　　国家级 口　市级 口　院级 口

教学院长签字：（公章）

年　月　日

学校意见：

是否同意立项：　　是　口　　否　口

项目级别：　　国家级口　市级口　院级口

负责人签字：（公章）

年　月　日