华中科技大学创新创业训练计划

项目中期检查表

项目编号 _____

项目名称 _“卷皮网”用户画像及精准推荐实战__

项目负责人_____孙嘉轩_____

所在院（系） _____管理学院_____

实施时间 _____2018.4 – 2019.4_____

联系电话 _____15332339493_____

填表时间 _____2018.10.11_____

华中科技大学教务处制

一、项目研究进展情况说明（条文列述）

1.与卷皮网进行对接，深入了解企业运营实际情况和与我们合作的潜在可行性。通过获取卷皮网网页数据（包括商品相关数据，销量及营销情况数据，用户评论数据等）以及卷皮网提供的内部数据库数据（包括规范化的企业运营数据，用户详细信息，浏览记录及订单转换率等），我们对卷皮网面向的用户群体以及运营情况进行研究与统计，运用机器学习的方法描述数据中所蕴涵的规律，并与企业进行交流，汇报研究情况，获得企业的反馈。

2.综合评价对卷皮网的调研与分析，小组成员决定在原计划的基础上进行一定程度的扩大，从数据的丰富程度，研究方法的多样和结果评估的多维度进行了扩充。使得我们可以从多个维度研究、评价和优化用户画像以及理解用户画像在不同场景下能够产生的实际作用。

3.全组成员深入学习统计学与机器学习在用户画像中的应用方法。并将自己的部分学习成果进行归纳总结，汇总成教程，并就自己在学习中的体会以及遇到的问题进行了总结。我们已经完成 python 语言基础教程，R 语言基础教程，对于用户画像的基本介绍，以及机器学习相关概念的介绍。后续我们将提供更加深入与独到的教程。

4.建立了项目组的官网，初步形成了之后教程的模式。通过 GitHub Pages 编写了官网，并根据模板编写了网页的后台代码，完成了界面 UI 设计。撰写了对于本教程的汇总，初步完成了教程大纲，上传了对于整体工作模式，教程思路，应用场景，案例分析等具体子模块的规划与样例。对相关内容框架进行了细分，在最大程度上确保了内容的完整性，丰富性和生动性，确保拥有初级编程敬仰的读者也可以流畅阅读教程，并且通过一些简单的案例教学进行上手实践。

5.进行了文本分析的实践应用尝试。第一种方法是通过 Twitter 及 Reddit 的企业 API 接口从服务器获取用户的发帖、评论、点赞数据，以及围绕某一主题所发表的各类文字、图片、表情等。通过文

本分析的程序包，爬取围绕某一关键词不同用户发表的贴文和评论，并根据其具体发帖内容进行分析，制作用户对于某一舆论热点事件的讨论的画像，对互联网上的舆情进行分析。第二种方法是爬取某些特定用户的关注人列表以及粉丝列表，通过大量的相关用户介绍来对该重点用户进行分析，以得出其在互联网上的影响力以及其发帖的重要性。以上两种方法均取得显著的结果。

6.上传一些简易版数据集并进行简单的分析，可供读者进行练习与研究。我们找到了不同行业的数据并进行了数据清理，标准化，特征提取等工作，使得数据能够更方便的供初级用户进行分析。我们也将一些分析的思路以及代码实现原理提供出来，给读者提供一定的启发。

二、已取得的阶段性成果（条文列述）

1.掌握用户画像作为一种新型商业分析手段的体系。研究了如何对用户所展现的社会属性，生活习惯，消费行为，意见倾向，情绪倾向等进行刻画。在单个用户刻画的基础上，尝试运用统计学的观念对用户进行分类和归纳，提炼出用户的群像，并对整个用户群进行特征刻画。最后，总结出了系统性的框架，讨论用户群像作为一种新型商业分析手段能够如何在实然层面对商业产生影响。讨论了在不同领域之间用户画像产生影响的区别与联系，如在新兴产业如互联网行业中，用户画像对用户的感知与影响体现在非常直观的层面，而对于传统行业如传统零售业，用户画像的影响或许较为间接，但同样可以对企业效益产生很大的推动作用。

2.掌握了构建用户画像的编程基础，了解机器学习的应用。团队成员分头对 Python 语言和 R 语言进行深入的学习与实践。作为机器学习应用中最常见，使用人数最多的两种语言，其配套的程序包非常丰富且功能强大。如 python 中的 scikit-learn,R 语言中的 quanteda，都是在实际操作中经常用到的程序包。熟练使用这些同行已经编写好的程序包，可以大大提高项目开发效率，同时降低编程的工作负担。

3.在项目开发上，团队成员学会熟练使用 GitHub 的版本库作为项目开发的合作方式。在 GitHub 上建立工作流、上传工作成果，提交修改意见并讨论，优化既有工作成果，团队合作变得十分简单高

效。同时，团队成员掌握了使用 GitHub Pages 进行网页开发的技巧。学会了修改设置文件，编辑 markdown 作为网页内容，修改网页 UI 格式。我们从海量的静态网页模板中找到最适合教程的模板并进行修改，使得网页界面最大程度的贴合我们的开发需求，给读者最好的阅读与互动体验。

3.结合已有技术，对“卷皮网” 用户画像构建进行尝试，并获得一定的结果。我们和在卷皮网工作的数据分析工程师取得了联系，获取后台的商品与部分交易信息数据。同时我们通过爬虫程序，从手机端的 APP 获得部分商品的用户评论数据。将两种数据结合起来，对用户评论与商品销量之间的关系进行研究。通过结果我们发现，在卷皮网上的服装类商品和食品类商品这两个板块，销量与评论数量存在较强的相关性，但在筛除了有大量重复的无效评论之后，数量关系不再显著。同时，好评率与商品销量存在较强的相关性。但对于卷皮网的用户画像研究出现了一些之前不曾预料的问题。一方面，卷皮网的价格较低，主要面向低端市场，因此低客单价的消费者是卷皮网的主力客户群体。而这部分人对于商品质量并无十分苛刻的要求，他们主要的关注点在于价格是否低廉。从评论的词频分析也可以看出，卷皮网的主要用户对于价格便宜这一特征的关注程度要显著高于其他商品特征，因此对于卷皮网用户画像的分析较为单一和有限。

4.提升了项目工作在广度和深度，并完成了基本教程框架的搭建。用户画像与机器学习是两个非常宽泛且内容丰富的领域，在近几年的学术界被广泛研究。且因为其学科交叉的特殊性，使得商科研究者与计算机科学研究者都对其投入了很高的关注度。在我们完成“卷皮网”项目时，我们发现仅仅通过单一的项目难以系统的实践并归纳总结出用户画像与机器学习不同方法的优缺点。且在查找资料的过程中，我们发现这两个领域前沿的研究，教程，论坛，程序包说明都是英文，这为中国研究者的研究增添了很大的麻烦。因此团队经过仔细的研究讨论，决定进一步扩大我们工作的范围，在广度和深度上更进一步，开发更多领域的应用场景，发掘更多方法在交叉融合中能够产生的新效果。同时将我们开发思路，研究日志，程序源代码，结果分析，优缺点分析等整理成通俗易懂的教程，不仅让自己的工作轨迹更加清晰明确，也为后来的初学者提供了学习，实践，讨论，提升的平台。在教程框架的初步设计中，我们一方面着眼于基本理论的介绍，从相关领域中找出最有代表性的理论，并将其写成通俗易懂的文章；一方面关注实践的重要性，因为对于编程最好的学习方式就是实践。所以我们将实践分成两个部分，一部分是我们较复杂项目的开发案例，将整个项目开发过

程呈现给读者，使得读者对项目整体开发思路和一些有代表性的方法有一定的了解。一部分是一些简单的入门级项目开发实例，我们通过一些简单的案例让初级研究者运用 python，R 以及程序包解决一些简单的问题，降低了上手的难度，让他们对项目开发有了最直观的感受并产生浓厚的兴趣。

三、目前存在问题

1.中文语义分析技术尚不成熟，对中文用户刻画的精确度较低。

2.优质数据来源的支持较为缺乏。企业对调用 API 有着严格的限制。

3.还未制定系统的教程推广方案，需要将相关受众群体转化为教程的固定读者。

四、经费使用情况说明

已获资助经费	0 元
已使用经费	1093.81 元
经费使用明细	1.相关书籍及学习资料： 《机器学习》61.6 《Python3 网络爬虫开发实战 Scrapy 数据分析处理》68.0 《深度学习入门之 PyTorch》63.8 《数据结构与算法分析：C 语言描述》25.6 《推荐系统实践》33.8 《用户画像：大数据时代的买家思维营销》 34.8 《推荐系统：技术、评估及高校算法》 105 《数据挖掘：概念与技术》 54.5 《R 语言数据分析与挖掘实战》 54.51 网课：Python 爬虫实战视频教程 159.2 大数据机器学习案例之推荐系统 168.0 合计 830.61 元 服务器及域名租赁：70 / 月 合计：210 3.打印费：55 元 总计 1093.81 元

五、下阶段研究计划及主要措施

1.进一步探讨用户画像在不同行业的应用场景，学习体会企业现成的用户画像在商业实践中的使用方法。对比不同方法在特定使用场景下的优劣。

2.从更多行业找到可供分析的数据并进行分析研究（例如，传统金融、互联网金融、购物网站、社交媒体等）

3.研究互联网上完善的用户画像开源项目，学习可以被借鉴的优势与特色，提炼优秀的方法整理成教程。

4.归纳总结 R/Python/MATLAB 等不同编程环境下的机器学习实现过程的区别与联系，充分发掘并利用不同软件的优势完成用户画像的建立与优化工作。

5.通过数据模拟，实际操作等方式研究如何量化用户画像对于企业经营利益产生的影响。

6.寻找简单的实际案例并提供解题思路以及解法，上传至 GitHub 供其他有相关领域开发意愿的初学者进行练习和交流。

7.将该网页应用到实际的课堂教学中进行实验，观察学生对用户画像与机器学习的兴趣以及对相关学习方法的接受程度。

项目负责人（签名）_____

年　月　日

六、项目指导教师意见

该项目组成员具备钻研能力和知难而上的精神，从零开始学习大数据分析，目前已经具备了对卷皮网进行用户画像所需要的数据基础和技术基础，完成的数据分析工作具有一定的实用性和推广应用价值。

项目组在顺利推进研究进度的同时，对自己提出了更高的要求，希望编写与机器学习和用户画像相关的入门教程，该教程的编写不仅是小组成员对项目研究成果的归纳总结，也对广大师生有学习参考价值，有助于提高后继学习者的学习效率。

鉴于小组的工作进展，同意通过中期检查，并推荐申报校级大创项目。

指导教师（签名）

年　月　日

七、项目负责人所在院（系）审核意见

审核人（签名）_____

_____年_____月_____日

注：请用 A4 纸，双面打印。