附件 1

创新训练

华中科技大学

大学生创新训练项目申报书

项 目 名 称 ： “卷皮网”用户画像及精准推荐实战

所属一级学科： 管理科学与工程

项 目 负 责 人： 孙嘉轩

专　　业　　： 物流管理

院　　系　　： 管理学院

申请资助经费： 3600

指 导 老 师 ： 张千帆

导师所在单位： 管理学院

实施起止时间： 2018.4—2019.4

填 表 时 间 ： 2018.4.8

华中科技大学教务处编制

一、人员基本信息（含项目负责人）

	姓　名	学号	入学时间	所在院(系)、专业	联系电话	项目分工
申请人或团队	孙嘉轩	U201615958	2016.9	管理学院、物流管理	15827510657	
	黄爽	U201715928	2017.9	管理学院、管理学创新实验班	15387118581	
	龙海文	U201616095	2016.9	管理学院、信息管理与信息系统	15827218773	
	杨岳浩	U201616098	2016.9	管理学院、信息管理与信息系统	13125198758	
	许艳	U201717035	2017.9	管理学院、管理学创新实验班	18571489296	
	团队名称：					

指导教师	姓　名	张千帆	年　龄	43	工作单位	华中科技大学管理学院
	职　称	教授	职　务		E-mail	qianfan_zhang@hust.edu.cn
	研究方向	企业信息化、社会化电子商务、数据分析			联系电话	13971397377

二、项目研究目的

本项目旨在通过对“卷皮网”商品层和用户的购买行为的数据分析，建立模型进行用户画像并实现精准推荐，为商家改善商品和服务并进行精准营销提供有效的模型基础。

“卷皮网”背景：“卷皮网”是以“平价电商”为定位的电商网站。其前身为“9 块邮”，是一家导购网站，但 2015 年它开始建立自己独立的电商平台，“9 块邮”变为其中一个版块。平台通过“弱品牌+非标品”的方式与其他电商巨头进行差异化竞争。其用户群体特征明显，主要为三四线城市年龄 18—35 岁的学生和中低收入人群，而且其中女性客户占比 70%以上。虽商品价格低廉，但其平均客单价在 50—60 元，用户复购率高。前期调研结果显示“卷皮网”的客户特征明显，其购物行为具有群体性，通过大数据分析有可能把这些具有群体性的购物行为具象，通过用户画像的方式体现出来。

三、项目研究内容

1. 数据收集与预处理：通过爬虫算法收集“卷皮网”网站上商品信息，用户评论，用户购买行为等维度的数据，并对用户评论及购买行为等非结构化数据进行分词及关键词提取；
2. 数据挖掘：分析商品数据、用户评论和用户购买行为数据，识别商品销售规律及用户购买习惯；
3. 训练模型：基于数据分析结果，构建、检验并评测模型，形成用户画像；
4. 应用推广：基于商品销售规律以及用户画像的研究结果，进行精准营销，为用户精准推荐商品。

四、国内、外研究现状和发展动态

1.用户画像描述产品用户群体行为特征，为商家提供多方位的用户信息，使商家了解、认知用户，改进产品或服务，为精准营销带来可能。例如，在酒店用户画像中，已有学者以在线评论数据为基础，从用户信息属性、酒店信息属性和用户评价信息属性三个维度构建用户画像模型的概念模型，并采用 Protégé工具建立本体来实现用户画像属性之间的关联，完成对酒店用户特征的完整刻画。还有学者对服装企业的用户构建了用户画像数据库，并基于 4C 理论从营销的角度构建了精准营销细分模型,并以三枪集团为背景进行案例分析，

研究结果表明以用户画像数据库为基础的精准营销细分模型能够重构消费者的需求、精准定位消费者群体，并能为服装企业实施精准营销提供科学的决策依据。本项目研究的用户不是实体企业的用户，具有非实名认证，用户的个人信息存在缺失的特点，研究难度更大，但也更具有挑战性。

2. 随着计算机技术的日益发展，在各个领域中所采集的数据集规模不断增大，特别是高维数据中存在的大量冗余和无关特征给机器学习带来了巨大的挑战。特征选择是为了解决高维度数据计算问题而衍生的，通过剔除冗余特征和无关特征，提高机器学习算法的泛化性能和运行效率。随着研究的深入，特征之间复杂的相互关系对机器学习算法的影响被逐渐地认识到，如何在特征选择过程中识别和保留具有交互关系的有益特征组合，是目前仍未很好解决的难题。Filter 特征选择算法能够辨别特征相互作用中冗余和依赖关系，进而选择出高度相关、内部依赖和低度冗余特征子集。针对基因表达数据在疾病诊断中的应用问题，提出了基于动态相关性分析的基因选择算法。此外，基于 Banzhaf 权利指数的特征评估及选择算法、基于 Shapley 值的特征选择算法优化方法和基于动态加权的特征选择算法，在公开测试数据集上的实验结果中均获得了良好的性能，达到了预期的效果和目的。以上算法为本项目的研究提供了理论基础。

参考文献

[1] W. Maass and T. Kowatsch, Semantic Technologies in Content Management Systems: Trends, Applications and Evaluations: Springer Science & Business Media, 2012.
[2] J. Vom Brocke, A. Simons, and A. Cleven, Towards a business process-oriented approach to enterprise content management: the ECMblueprinting framework.Information Systems and e-Business Management, 2011,9:475- 496.
[3] Schafer J. B., Konstan I., Electronic Commerce Recommender Applications, Journal of Data Mining and Knowledge Discovery, 2001,5(1-2):115-152.
[4]S.S.Ahila and K.L.Shunmuganathan,Role of Agent Technology in Web Usage Mining:Homomorphic Encryption Based Recommendation for E-commerce Applications, Wireless Personal Communications,2016,87(2):499-512
[5]Liu, Ding-Yu, Chia-Sui Wang, and Kuei-Shu Hsu. Beacon applications in information services. Advanced Materials for Science and Engineering (ICAMSE), International Conference on. IEEE, 2016.
[6]Gupta, Rajan, and Chaitanya Pathak. A Machine Learning Framework for Predicting Purchase by Online Customers based on Dynamic Pricing. Procedia Computer Science 2014,(36): 599-605.
[7]Kerin, R. A., Jain, A., & Howard, D. J. Store shopping experience and consumer price-quality-value perceptions. Journal of Retailing, 1992,68(4), 376.

五、研究路线及解决的主要问题

研究路线：包括数据收集和预处理，模型生成，评测应用三个阶段。

1.数据收集和预处理

（1） 数据收集：利用爬虫技术收集商品基本信息和用户评论信息，对用户浏览及购买行为进行持续性、大样本的数据收集。

（2） 预处理：先去除无效评论，再对有效评论进行分词处理，构建特征词典，从而分析语意和情感倾向。

2.模型生成

（1） 训练数据：采用样本采样，样本标注等方法训练数据。

（2） 特征提取：采用主题分析，分词，特征扩展，关键词提取等方式获得标签，构建模型。

（3） 模型训练：采用分类，聚类，分布式计算等技术训练模型。

3. 评测应用
模型评测：对模型进行测试和验证，根据反馈调整和优化模型。

拟解决的主要问题：
1. 用户行为等动态信息的获取。
2. 评论的语意识别与情感分析。
3. 对设置用户的标签的准确性和模型算法的选择。
4. 用户画像后与分类的商品的对应。

六、项目创新及特色

项目创新特色概述
1.选择网站的创新：“卷皮网”的定位是“平价电商”，其商品和用户都具有明显的特征，适合于通过模型来进行分析与类聚。并且网站并没有成熟的精准推荐系统，我们的研究可以对网站的发展起到积极意义。
2.文本处理方法创新：构建特征词典、情感词典，对有效的用户评论做多元处理。
3.方法的创新：大数据的获取与处理与电商网站的营销策略相结合，利用构建模型的方向做精准推荐。

七、项目综述

（前期预研基础、自身具备的知识条件、项目实施的科学性、创新性及技术可行性等）
1. **前期预研基础**：据统计，2017 年底中国网购用户规模达 4.8 亿 人，中产阶级电子商务消费群体的崛起使得企业由以“产 品”为中心向以“用户”为中心转换，如何更好的了解 用户需求，以推动用户购买决策的形成成为企业急需解 决的问题。用户画像技术本质是指企业在大数据环境下， 洞察用户信息，全面准确的刻画用户的维度与属性，使 企业能深入研究用户特征与行为以更加精准的掌握用户 需求，从而更好的服务于用户。
2. **自身具备的知识条件**：参与人员皆具有良好的编程基础，同时掌握了网络爬虫获取数据、清洗数据、处理数据以及数据挖掘的原理，均能使用 C++以及 Python 进行编程，另对电子商务网站用户的购买决策过程了解透彻。
3. **项目实施的科学性**：用户画像作为大数据的根基，它完美地抽象出一个用户的信息全貌，为进一步精准、快速地分析用户行为习惯、消费习惯等重要信息，提供了足够的数据基础，奠定了大数据时代的基石。以往关于电商平台的研究基本都是从平台的 度分析它的盈利模式、商务模式，极少从用户角度分析平台的营销活动进而达到盈利目的。本项目选择从用户画像的角度，重点梳理和构建了用户画像的体系及精准营销实施的理论框架， 帮助电商平台了解用户真实的特点及行为轨迹，完善产品，优化用户体验，并且通过用户细分，锁定目标群体，对目标群体实施不同营销活动，进一步提升满意度。
4. **项目的技术可行性**：本项目基于用户基础数据，包括静态数据和动态数据（主要为网络行为数据、平台内用户行为数据、用户内容偏好数据、用户交易数据这四类），通过文本挖掘、自然语言处理、机器学习、预测算法、聚类算法等方法对收集到的用户基础数据进行行为建模，在此过程中抽象出用户的标签，从而实现下一步的精准营销。

八、项目实施方案

项目方案（进程安排等）
1. 通过阅读文献和书籍，学习 Python 爬虫和机器学习相关算法，如支持向量机（SVM）、人工神经网络（NN）等，了解算法的详细内容，并找到原始代码。（2018 年 4 月——5 月）
2. 开始使用爬虫对网站商品信息、用户行为信息及用户评论进行爬取。（2018 年 5 月——7 月）
3. 学习一些数据处理软件（SPSS、MATLAB 等）并实时预处理获得的数据，并学习无结构化文本的语意分析算法，对用户评论进行初步处理。（2018 年 5 月——7 月）
4. 开始结合获取的用户行为信息和用户评论给用户分类、贴标签，为下一步建立模型做准备。（2018 年 7 月——2018 年 9 月）
5. 通过不断获取，综合用户信息和商品信息尝试建立精准推荐的模型。（2018 年 9 月——12 月）

6. 继续收集用户行为信息，对模型进行验证、优化（2019 年 1 月——2 月）
7. 通过后期数据对模型进行检验，并对项目成果进行总结。（2019 年 2 月——4 月）

九、项目预期成果

1.通过对部分电商网站商品信息以及用户购买行为和评价的收集和分析，对商品进行多维度的分类，进一步确定商品的目标人群以及市场投放策略。
2.根据用户购买行为以及评论，明确影响用户行为产生的因素，生成标签，构建完整的用户画像，生成一套精准推荐算法。
3.结合用户层与商品层的数据，综合得出提高电商企业运营效率，营销水平的方案，为企业精准营销提供决策依据，增加企业利润，减少不必要的成本。

十、经费预算

项目经费预算（主要内容为印刷费、调研差旅费、材料费、测试及加工费、图书资料费、论文版面费等）
文印费：800 元
调研费：1000 元
图书资料费：800 元
数据获取费：1000 元
合计：3600 元

十一、审批情况

指导教师意见	大数据分析技术的应用前景广阔，值得师生积极学习和探索。本项目组的几位成员具有很强的钻研精神和自主学习能力，在学院没有开始相关课程的情况下，已经自学了 Python 和 SPSS，并对卷皮网的网上数据进行了初步的分析，预研工作证明了本项目的选题正确，团队成员也表现出了良好的团队意识。本项目的研究目标明确，研究思路清晰，研究方法有难度但是通过努力可以掌握。鉴于对团队成员的信任以及对项目的认可，本人特推荐本项目申报华中科技大学大学生创新训练项目。 签名： 年　月　日
院系意见	院（系）（章）签名： 年　月　日
学校意见	

	教务处（章）签名： 年　月　日