武漢大學

研究生学位论文开题报告登记表

学　　院：计算机学院

专　　业：软件工程

学　　号：**2015212163317**

姓　　名：王 焰

导师姓名：郑 宏

导师职称：教 授

武汉大学关于研究生学位论文开题报告的有关规定

根据《中华人民共和国学位条例》及其《暂行实施办法》和《武汉大学学位授予工作细则》的精神，为做好研究生学位论文的开题报告，保证学位论文质量，特作如下规定：

第一条 学位论文开题报告是研究生写作论文的必经过程，所有研究生（含：博士生、硕士生）在修完学位课程，写作学位论文之间都必须作开题报告。

第二条 开题报告主要检验研究生对专业知识的独立驾驭能力和研究能力，考察写作论文准备工作是否深入细致，包括选题是否恰当，资料占有是否翔实、全面，对国内外的研究现状是否了解，本人的研究是否具有开拓、创新性。

第三条 学位论文开题报告前，研究生必须根据专业培养目标，结合导师、教研室（或研究室）所承担的国家、省部委等有关部门下达的研究项目或课题以及本人的研究特长，与导师协商，确定选题，广泛查阅文献，深入调研，收集资料，制定学术研究方案，在此基础上撰写开题报告。

第四条 研究生进行开题报告，必须提交“开题报告”的书面材料，内容包括：（1）论文选题的理由或意义；（2）国内外关于该课题的研究现状及趋势；（3）本人的研究计划，包括研究目标、内容、拟突破的难题或攻克的难关、自己的创新或特色、实验方案或写作计划等；（4）主要参考文献目录。开题报告的书面材料不得少于 3000 字。

第五条 研究生进行学位论文开题报告要向导师提出申请，申请获准后，博士生在博士生指导小组范围内作开题报告，硕士生在导师所在教研室或教学小组作开题报告。参加开题报告的教师，包括导师在内，一般不得少于 3 人。无论博士生还是硕士生，在作开题报告时，本学科专业的研究生一般必须参加，跨学科或相近专业的研究生亦可旁听。

第六条 参加研究生学位论文开题报告的教师应当对开题报告进行评议，主要评议论文的选题是否恰当，研究设想是否合理、可行，研究内容与方法是否具有开拓性、创新性，研究生是否可以开始进行论文写作等。评议结果分“合格”与“不合格”二种。评议结束后，由研究生指导教师在《研究生学位论文开题报告登记表》“评语”栏中填写评语。学位论文开题报告通过后，研究生方可进行论文撰写工作。

第七条 开题报告结束后，研究生应将登记表复印一份连同登记表原件和开题报告等一并交所在院、系研究生干事将登记表复印件加盖公章后报送研究生院培养教育处，其他材料留存院、系查备查。研究生院培养教育处将不定期抽查研究生开题报告材料。

第八条 本规定自 2008 年级研究生开始实行。

第九条 本规定由研究生培养教育处负责解释。

武 汉 大 学 研 究 生 院

研究生学位论文开题报告表

姓　　名	王焰	院、系（所）	计算机学院
学 科 专 业	软件工程	攻读学位	工程硕士
研 究 方 向	软件工程	指导教师	郑宏
拟定学位论文题目：　　基于大数据分析的广告精准投放研究			
参加开题报告教师人数	4	参加旁听学生人数	10

	姓　　名	职　　称	所 在 工 作 单 位
开题报告组成人员	郑宏	教授	武汉大学电子信息学院
	杨文	教授	武汉大学电子信息学院
	孙涛	教授	武汉大学电子信息学院
	王文伟	副教授	武汉大约电子信息学院

研究生开题报告及指导教师所提问题回答的内容记录：

一、论文名称、选题依据

1、论文名称：基于大数据分析的广告精准投放研究

2、选题依据：

随着移动终端数量的迅速增长，移动端广告已经成为了互联网广告的主导部分，占据了绝大部分的市场份额[1]。与此同时，移动端媒体具有本身的特性，如移动化、精细化和个性化等，这给移动端广告精确化投放提供了基础，意味着移动端广告具备个性化推送的能力。移动端广告精细化投放的优势是能够提高广告与消费者消费行为习惯的匹配度，促进用户消费，从而增加广告的商业收益[2]。然而，目前的广告投放绝大部分都是粗放型的投放方式，不具备个性化和精细化的特征。少量的个性化广告投放是基于内容的推荐，仅仅根据用户当前页面的关键词，然后匹配相关的广告，并没有考虑用户自身的兴趣从而进行个性化推荐。

随着信息技术的不断发展，互联网产生的咨询以指数级的速率增长[3]。在近乎无穷的信息中大量的数据与用户的兴趣并无太大的相关性，这意味着存在着信息冗余的问题，解决信息过载的问题已经成为了信息科学领域的关键问题。具体到实际的问题当中，上述问题表现为如何使用户高效地获取有用的信息，对于信息服务提供商来说，面临的问题是如何把与用户相关性高的信息推送给用户，从而提高信息服务的收益。

互联网广告是一种信息服务，在互联网广告投放的过程中也需要面型信息过载的问题。在解决互联网广告投放的问题，要考虑互联网广告本身的特性。从而实现精确投放[4]。

二、本课题国内外研究现状及发展趋势

近年来互联网广告行业发展迅速，已经成为互联网公司盈利最高的部门之一，显示出了良好的前景。我国的互联网广告规模在逐渐增长，到 2014 年，成为了全球第二大的互联网广告市场。在 2001-2015 年期间，我国广告市场的增长率达到了 14.32%。而随着移动互联网的发展，移动终端的普及，已经移动应用如社交媒体、视频网站和电商网站的成熟，互联网广告的形式日趋多样化。

目前国外在精准广告投放领域比较突出的成就有 Google 的 Adsense，早在 2003 年，谷歌公司就开始将自己的广告商网络提供给第三方使用，如果用户通过 Adsense 点击了广告，那么广告商将根据点击情况向谷歌付费；微软也推出了 adCenter，通过跟踪用户的消费行为[17]，选取有价值的关键词和目标网站，更好的为广告主寻找潜在目标客户。

准确的广告可以基于广告商和广告内容，在 Cina Apple 的数字媒体交易平台中，选择目标用户和地区，使用汉字、图片或视频，准确的广告给用户。准确的广告具有以下优点：（1）准确的广告提高了广告的准确率和命中率。准确的广告需要准确的市场细分。数据挖掘技术用于收集、处理、保存和分析用户信息，并将目标广告准确地传递给目标用户。因此，基于数据挖掘技术的精确广告提高了广告的准确率和命中率，节约了广告成本，满足了对目标用户集中有限资源的经营理念，提高了企业的投资回报率。（2）精准广告提高了电子商务服务水平。为了实现准确的广告投放，电子商务必须以用户的需求为出发点和终点，分析用户需求的变化，尽可能地满足用户的需求。同时，电子商务也要求用户提供最大的节约和便利，减少用户的消费渠道，选择合适的物流服务。

三、研究目标

针对以上问题，本文拟通过对楼盘数据的采集与挖掘，采用最新的机器学习技术，创建关联规则，对不同层次受众群体进行聚类分析，建立广告投放模型，并通过不同领域广告投放商对广告位的选择进行协同过滤以及组合分析，实现广告投放商向群众的精准广告投放，在广告商与受众群体之间实现最优匹配。除此之外，我们将基于已获得的数据建立

特定领域的知识库，进一步挖掘用户群体与广告的深层次联系

四、课题研究的主要内容

摘 要

Abstract

第一章 绪论

1.1 选题背景和意义

1.2 相关技术的发展和研究现状

1.3 研究内容

1.4 论文结构

第二章 相关理论与关键技术

2.1 引言

2.2 机器学习和数据挖掘

2.3 用户行为

2.3.1 用户行为的概念

2.3.2 用户行为的分类

2.4 协同过滤

2.4.1 推荐系统

2.4.2 相似度的计算

2.4.3 相似邻居的计算

2.5 本章小结

第三章 地缘特征数据挖掘和分析

3.1 引言

3.2 数据爬取

3.2.1 网络爬虫

3.2.2 爬虫流程

3.2.3 模块细解

3.3 数据挖掘方法

3.4 数据挖掘结果分析

3.5 用户画像与广告商画像

3.6 本章小结

第四章 广告精确投放算法与冷启动

4.1 引言

4.2 数据标引流程

4.2.1 算法概述

4.2.2 数据采用

4.2.3 数据量化

4.2.4 总体结构

4.2.5 计算结果

4.3 广告投放推荐模型

4.3.1 核心问题解决方案与算法

4.3.2 基于用户的协同过滤和基于商品的协同过滤

4.4 冷启动系统

4.4.1 利用用户历史数据

4.4.2 利用用户兴趣偏好

4.4.3 通过选项采集初始信息

4.4.4 广告系统冷启动
4.5 本章小结
第五章 系统的实现与实验评估
5.1 引言
5.2 需求分析和系统设计
5.2.1 市场需求分析
5.2.2 用户需求分析
5.3 系统架构
5.3.1 开发环境介绍
5.3.2 数据库设计
5.3.3 系统原型界面展示
5.4 实验结果评估
第六章 总结与展望
6.1 总结
6.2 展望
参考文献
致　　谢

五、拟解决的问题

冷启动数据模块主要作用是数据收集、清晰和分析。数据采集的工作主要靠 Web 爬虫，从各大网站上爬取数据，比如从搜房网、房天下、链家网、地产网等房产数据，这些房产数据主要包括地理位置信息、房价信息、交通信息、户型信息、建筑年代、配套属性等信息。同时要对数据进行清洗，失效数据需要通过正则匹配等方式进行清洗。然后对缺失值进行评估，重点是对数据进行特征提取。广告位推荐系统主要工作是设计定价模型，根据已知小区房价水平，似然评估其消费水平，不要求绝对准确，用作推荐系统冷启动数据基础，根据广告位所在小区房价水平和周边商圈密集程度，为广告位价格做初始评估。人物画像系统的主要作用是分析用户行为，根据房价信息和商圈信息对用户的消费能力进行评价

六、拟采用的研究方法和实验方案

1.算法综述

地段价值的高低也是评估购买力的一个有效标准。面对大笔金额差异，处于上层精英社会的人群相比普通老百姓大概率住在更加优质的地段。

地段价值进行评估在当前时代极具价值，比如买房，门面选址等等，决定地段价值的因素也越来越显著可寻。对于选址来说，在经济允许的条件下，更专注关于地段相关的特征选取，通过对地段周边各类设施配套水平，交通，商场等进行学习。

本算法则更偏向于高效投资，这也是对大多数人而言更关注的地方，地段的实际价值不容易准确评估量化，但地段与地段之间的好坏比较是显著的，在不考虑政策相关因素时，通过已经得到的数据将地段划分为 5~10 类，从低价逐步到高价的类型。而后通过深度学习训练得到每类地段周边数据模式，进而找出最具购买力的地段。

算法的总体逻辑是将地段根据房价划分为 5~10 类，将地段周边的数据量化分层输入深度学习模型进行训练，进而找出影响地段价值变化的关键因素。然后将所有已知城市数据量化输入，进而找出最具投资价值的地段。

2.数据采用

首先为了对地段价值进行评估，需要了解其周边各类设施配套水平，如交通，商场，学校等等，这些数据可以通过路网，POI 结果中得到。其次数据获取后需要被二次区分，优

质商场或者学校对于地段的影响远大于普通商场与学校[10]。然后数据需要被合理量化，更标准一致的数据集可以更好的影响最终训练结果[11]。

3.数据量化

（1）以被标引的数据为中心，将周边 2 公里范围离散化离散化数据可能会损失部分有价值的细粒度数据，但同时带来的好处是可以直接准确的评估数据的影响程度。比如商场 POI 对结果的影响，甚至商场 POI 的具体面积，占地结构对结果的影响。

（2）启动前将对各种不同的地段数据集进行标引，确定数据集属于某种具体的分类，通过已有的房价信息将地段划分为 1、2、3…等多个等级

（3）将多元数据分别映射至各个离散区域中，形成量化数据，利用 0-1 矩阵对图像进行二值化处理，得到一个 0-1 矩阵，通过这个方式可以将河流的图示转化为可以用于计算的矩阵，有利于建立量化评价体系和算法分析。

利用 0-1 矩阵对图像进行二值化处理，得到一个 0-1 矩阵，通过这个方式可以将公交线路的转化为可以用于计算的矩阵，有利于建立量化评价体系和算法分析。

（4）得到标记样本的数据集

（5）得到拥有一组样本的训练集

1. 训练过程：将 80%已标记地段样本的量化数据分离到 C1 状态（分离出路网，地理信息，POI，人流等量化特征），由多层感知器提取其最显著特征到 Sn，在 NN 阶段将特征向量重新连接，计算调整特征输出权重直到标记样本用完或趋近收敛。
2. 验证过程：将 20%已标记地段样本的量化数据作为输入，通过 NN 计算得到 Label，计算准确率与召回率。
3. 应用过程：将城市所有未标记数据作为输入，计算得到城市地段价值 Label。

算法的整体步骤如下：

1. 输入层：输入为带标签样本量化后的数据集（分层表示路网，地理信息，POI，人流等）
2. C1：将量化数据集拆分出有效特征（C1 在处理阶段已量化，将每层拆分）
3. S1：二次滤波提取有效特征
4. Sn：n 次卷积后的结果数据集
5. NN：映射为线性向量后的神经网络分类
6. Label：最终决定分类的 Label

七、研究计划，进展和预期答辩时间

- 2017 年 11 月 01 日到 2017 年 11 月 25 日，确定论文选题
- 2017 年 11 月 25 日到 2017 年 11 月 30 日，撰写开题报告和开题
- 2018 年 3 月 17 日到 2018 年 4 月 10 日，搜集资料和阅读文献，完成论文提纲
- 2018 年 4 月 11 日到 2018 年 4 月 15 日，深入研究形成初稿
- 2018 年 4 月 15 日到 2018 年 5 月 10 日，论文修改、定稿和答辩

八、主要参考文献：

[1]. 刘朋，林泓，高德威.基于内容和链接分析的主题爬虫策略[J].计算机与数字工程，2009,37(1)：22-24，80.

[2]. 陈晨.基于主题爬虫的个性化搜索引擎技术研究[J].科技信息，2010,(31):87.

[3]. Punnawat Tadapak, Thanaphon Suebchua, Arnon Rungsawang. A Machine Learning based Language Specific Web Site Crawler[A].13th International Conference on Network-Based Information Systems[D]. New York:IEEE,2010.155-161.

[4]. 郑志高，刘庆圣，陈立彬.基于主题网络爬虫的网络学习资源收集平台的设计[J].中国教育信息化，2010.01：36-38.

[5]. 张晓阳.基于 cookie 的精准广告投放技术及其法律边界刍议[J].电子知识产权，2015，9：

[6]. 程龙龙.基于 LDA 的行为定向广告投放算法研究[D].中国辽宁.辽宁大学.2014.

[7]. 何家瑞.广告精准化的策划与投放策略[J].企业改革与管理，2013，4:74-75.

[8]. 温爱华，郑艳娟.数据挖掘在保险客户关系管理中的应用[J].中国商贸，2009，62-63.

[9]. 商锦博.探索数据挖掘在保险公司中的应用[J]. 商场现代化，2007，7：163-164.

[10]. 张强，吕军.基于 Agent 和数据挖掘的分布式信息审计平台[J],2006,4:141-146.

[11]. 梅强，张冬荣.数据挖掘在保险分析中的应用[J].计算机工程，2004，12：571-573.

[12]. 刘梦超，陈荣，贺祥.数据挖掘在用户上网行为分析中的应用研究[J].电脑知识与技术，2012，11：7409-7412.

[13].王昭，数据挖掘在电子政务中的应用[J]河北联合大学学报.2013，4：78-80.

[14]. 刘玉宏.数据挖掘在保险客户关系管理中的应用[J].信息与电脑，2015，8：44-45.

[15]. 刘丽.基于数据仓库的保险管理系统的设计与实现[J].微机发展，2004，7：55-58.

[16]. 李琴.广告的行为定向投放技术研究[D].中国广东.广东工业大学.2014.

[17]. 柴源.网络用户行为分析及其预测技术研究[D].中国北京.北京邮电大学.2013.

开题报告记录人签名：

年　　月　　日

<table>
<tr><td colspan="3">指导教师意见：

指导教师签名：
年　　月　　日</td></tr>
<tr><td rowspan="2">开
题
报
告
评
语</td><td>评议结果</td><td></td></tr>
<tr><td colspan="2">参加开题报告的教师（3~5 人）签名：

年　　月　　日</td></tr>
</table>

注：评议结果分“合格”或“不合格”。