各位老师，下午好，我的名字是王焰，我的论文题目是基于大数据分析的广告精准投放研究，论文的指导老师是郑宏教授。

下面我主要通过五个方面介绍我的论文工作 首先是介绍研究背景和相关工作，然后是研究框架和系统的设计与实验，最后给出本文的总结与展望。

随着经济的进步与发展，有效的广告投放越来越成为企业发展的重要部分。对于公司来说，若想要用最小的广告投入实现最大的信息传递效果、用户数量增加、产品知名度的提高，与客户之间进行精准的信息交互显得特别重要。目前城市中所用的广告位招商，仍采用过去广告投放商自主选择（投放位置），然后广告制作商进行广告印制，再然后由物业进行定点投放广告的传统方式。这种广告招商方式存在很明显的问题：

（1）商家选择广告投放位置具有盲目性，位置没有经过详细调研，这种行为严重降低了广告投放的有效性。

（2）没有构建足够简单的广告发布网络，商家的广告发布流程复杂化。

（3）商家投放广告的资金分配没有建立适当的经济学模型，对投资策略进行合理化评估，从而得到用户回馈与投入资金的最优匹配。

当我们在淘宝浏览商品或购买商品后，淘宝总是能够给我们推荐称心如意商品。

亚马逊会根据我们的图书购买记录，为我们推荐下一本挚爱读物。

今日头条等咨询类的APP，强调你关心的才是头条。

他们都是根据用户的行为来做精准的推荐，我们能否在广告领域做精准推荐呢？

国外的谷歌和微软也有相关的研究，他们都可以通过跟踪用户的消费行为，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。

为了解决这一问题，如图所示，我们为广告商设计了一个推荐系统，本推荐系统共有三部分构成，即冷启动数据系统、广告位推荐系统、人物画像系统。

冷启动数据模块主要作用是数据收集、清洗和分析。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、地产网等房产商获取数据，这些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，并且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。

为了解决这一问题，如图所示，我们为广告商设计了一个推荐系统，本推荐系统共有三部分构成，即冷启动数据系统、广告位推荐系统、人物画像系统。

冷启动数据模块主要作用是数据收集、清洗和分析。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、地产网等房产商获取数据，这些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，并且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。

我们通过爬虫收集数据，通用网络爬虫的主程序主要由调度器，解析器和资源库三部分组成。调度器主要负责给主程序中的各个爬虫线程分配工作任务。调度器是网络爬虫的中央控制器，它根据系统传过来的URL，分配线程，启动此线程以调用爬虫爬取网页。解析器负责下载网页，解析页面，处理网页的内容，爬虫的基本工作是由解析器完成的。资源库用于存储下载的网页等资源。

我们爬取到的数据其实是杂乱无章的，包含非常多的无关信息。为了获取我们想要的信息，采用图中所示的方式，对数据进行清洗和分析。主要通过网页解析器来完成。如图所示，网页解析器主要经过词法分析和语法分析。对由爬取模块爬取的每个网页，词法分析器对其进行分词，保存为一个个的词法单元。这样来得到想要的信息。

目前已经完成了对搜房网、房天下、安居客、焦点房地产网、365地产家居等 房产网站的数据爬取。对于爬取网站上楼盘信息进行了成分提取与多态化分析，为项目的开展奠定了良好的基础。

除了前面我们爬取到的地价和商圈信息之外，我们还结合了交通、河流等地理信息，通过对他们进行二值化处理，转化为矩阵，有利于特征提取。

（一）算法的整体结构如图所示。

总体结构为：

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，在神经网络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。

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。

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。

（二）整体步骤为：

输入层：输入为带标签样本量化后的数据集，具体分层表示路网，地理信息，POI，人流等。

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）

S1：二次滤波提取有效特征

Sn：n次**卷积**后的结果数据集

NN：映射为线性向量后的神经网络分类

Label：最终决定分类的Label

从图中可以看出，武汉的地段价值指标呈现出沿江为中心的趋势，不同的地段价值 由图像中柱形图的高低来表示。柱形图越高，表示该地的地段价值越高，同时也表示该地区的购买能力越高，广告的投放价值也越高，不同的颜色代表了适合不同类型的广告投放建议。

用户行为是指网络上的用户操作：一般包括：用户经常浏览的网站、在浏览器或者其他搜索引擎中的关键字；用户打开网页的时间段，浏览记录，浏览次数，浏览时长和入口形式等。要从用户的行为和偏好中发现规律，并基于此给予推荐，收集用户的偏好信息，非常关键。得到用户行为之后，我们就可以进行用户画像。

左边图上给出的是用户的偏好，右边是根据用户的偏好完成的用户画像。

得到用户的行为信息之后，我们就可以通过协调过滤对其进行推荐。首先，我们可以通过相似度计算，把用户进行分类。对于相同类型的广告商，推荐相似的广告方案，也可以对相似的普通用户，进行广告推荐。

系统模块如图所示。主要分为三层，首先是客户层，也就是展现在用户面前的界面，这一层使用浏览器或微信来展示，通过向后台发起HTTP请求，来获取相关数据并展示，技术上使用Vue 来驱动视图的渲染。

其次为服务层，主要用来响应客户层的请求，处理数据和分析数据。这一层包括：访问控制层，业务逻辑层，数据控制层三部分组成。访问控制层使用Spring MVC Controller来处理与客户层的交互，业务逻辑层使用Spring Service来完成相关业务的实现，包括用户账户、广告位浏览、发布、购买、记录的逻辑操作，数据控制层使用Spring Dao控制管理相关数据，并与数据层进行关联。

最后是数据层，主要负责数据库驱动管理，包括数据驱动层和数据存储层。数据驱动层用于响应服务层的请求，并通过相关驱动或协议访问数据库或静态文件存储中的信息。

这个是系统选用的云服务器配置和相关环境介绍。

这是用户界面，系统可以根据广告商的需求来推荐最优的广告投放方案，并且分析广告受众的年龄组成、消费层次、购买能力等。

现在让我们再次测试原型系统的有效性。由于原型系统尚未投入运行，该系统仍在测试中。因此，暂时使用投资者旗帜下的某个保险业务网站进行广告初步测试，分析用户的行为，并检查原型系统准确广告的有效性。该平台是一个保险业务网站，已经运行多年，拥有庞大的用户行为历史数据。该平台不仅是使用该系统进行广告的广告商，还是广告平台的提供商。我们来看看衡量广告系统有效性的点击率。通过这个系统，我们可以知道广告的数量，点击量等信息，从而获得广告的点击率。我们测试传统营销和精准营销，结果显示在表格中。

通过以上结果我们发现，执行精准广告营销之后，虽广告显示次数下降，但被点击的次数和点击率都提高了。总之，精准广告营销比传统广告营销提高很多。

然后我们将上面表的结果绘成了图，可以看到精准营销点击率明显提高。

本文的创新点在于

（1）运用大数据分析的方法，实现了广告商广告的高效投放，减少投放资金，提高投放效果，极具市场价值。

（2）通过机器学习建立广告投放的资金分配模型，对于投资策略进行合理化的评估，降低了广告投放低回馈的风险。

（3）通过对受众群体的特征分析与广告商群体的特征分析，建立完整的信息模型，完成“用户画像”与“商家画像”。并基于已获得的数据建立动态立体的知识库，通过进一步研究挖掘用户群体与广告的深层次联系。

还有需要该改的地方：

1.协同过滤算法精度有待提高，需要结合其他算法，以提高推荐的精度

2.用户行为数据集采集不够全面

3.冷启动系统虽然可以给系统提供推荐的基础，但是数据的更新和同部同样非常重要