**基于机器学习和大数据分析的广告投放精准推荐系统**

**摘要**

随着移动设备的普及程度越来越高，移动广告占据的市场份额也越来越大。相比于传统互联网，移动媒体自身具有移动化、碎片化、个性化等特点，这就需要移动广告的投放必须朝着精准个性化的方向发展。只有改变传统广告粗放型的投放方式，针对不同用户迥异的兴趣进行个性化的广告投放，才能将广告投放转化为用户的消费行为，使广告投放商和广告主都能得到良好的商业回报。但是，现有的个性化广告推荐大都是基于内容的推荐，先提取用户当前页面的关键词，然后投放与之匹配的广告，并没有考虑用户自身的兴趣。

本项目采用最新的机器学习技术，通过对房地产楼盘、地域及周围商圈密集程度等大数据的采集与挖掘，创建关联规则，聚类分析不同层次受众群体以建立广告投放模型。另外，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，实现广告投放商向用户的精准投放。本项目将基于已获得的大数据建立特定领域的知识库，进一步挖掘用户群体与广告之间的深层次联系。

**Abstract**

# 第一章 绪论

## 1.1 选题背景和意义

我国的房地产行业目前存在巨大的泡沫，一方面房价在较危险的高位运行，另一方面却存在大量的房屋空置，房地产的“去库存”是国民经济的重大战略需求。实现房地产商与客户之间信息的精准交互是解决这一问题的关键。对于其他需要进行广告投放的商业公司，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、产品知名度的提高），与客户之间进行精准的信息交互依然是实现的关键。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）→广告制作商进行广告印制→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，大部分广告

（2） 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。

（3） 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。

（4） 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。

近几年来，信息技术和互联网发展势头迅猛，各种各样的资讯以几何倍的速度增长[3]。在完成原始的积累以后，互联网上的信息早就多到普通人一生也看不完的地步。面对如此庞大数量级的信息，用户如何高效的从中发现对自己有用的或者喜欢的信息，变成了一个难题。对于信息服务商来说，如何把自己想要推广的信息针对性的推送给相关用户，从而达到利益最大化，这也是个难题。

互联网广告的发展也面临同样问题，特别是在移动互联网的快速发展的当下，传统互联网占据的市场比例不断缩小。在这种大的环境下，如果想要抢占移动互联网的市场同时避免因广告的不合理投放造成的用户浏览和使用失衡，就要提高广告投放的质量。移动广告相比于传统互联网广告，主要呈现如下几个特点：精准性：移动广告在精准性方面有着先天的优势，它可以依据用户的实际情况和实情实景将广告推送到用户的移动客户端，实现精准投放[4]。

即时性：用户的移动设备一般随身携带，随时都能打开移动客户端访问，这是不受时间地点限制的，广告的投放也可以是全天候的，广告信息能及时有效的到达用户端口。

互动性：移动设备的发展给了交互方式更多的可能性，例如观看视频或者玩

HTML5游戏。随着用户与设备发生互动的时间积累，广告投放商就能获得更多的

用户需求。

扩散性：移动广告具有很强的再传播性。如果用户在移动端看到特别感兴趣的广告，就可以使用微信、微博等 App 自带的转发功能让周围的人也看到同样的广告[1][5]。

可追踪：一个移动设备背后的用户几乎固定不变，也就是说设备与用户是 1:1的对应关系，可以方便的对用户数量进行精确统计。同时移动设备自带的定位功能，扩充了用户数据可收集的维度。

用户使用移动端的次数变得更多，时间更短，投放空间更小，这些变化都促使广告投放商必须改变传统粗放型的广告投放方式，针对不同用户之间个体差异提供精准化的广告投放。因此有必要进一步研究如何建立用户的兴趣模型，了解用户的需求。如何在众多的广告资源中找到与用户的兴趣相符或高度相似的广告，无论是对于提升用户体验，还是提高客户端流量与广告投放商的收益，都显得非常有必要。

## 1.2 相关技术的发展和研究现状

## 1.3 研究内容和论文结构

针对以上问题，本项目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。本文的创新点在于：

（1） 打破了目前广告位投放市场仍保留的传统状态，实现高效的管理机制与投放信息推荐系统，实现房地产商与客户的精准对接，进而实现“去库存化”，符合我国经济发展的战略需求。

（2） 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。

（3）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。

（4）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。

本文共分为6个章节，每个章节的主要内容和组织结构如下：

第一章：绪论。概括的介绍了移动广告的发展状况，研究广告推荐技术的意义。同时也简要的阐述了广告推荐技术的发展。

第二章：介绍用户画像和个性化推荐的相关技术。这其中包括用户画像在大数据时代的重要作用，用户画像的流程等。接着阐述了几种常见的推荐算法，同时介绍了评价推荐系统好坏的标准与推荐系统面临的问题。

第三章：详细论述了用户画像的方法，于此同时针对建模中使用的 VSM 模型提出了改进。通过模拟仿真实验，论证了改进后算法的有效性，以及在文本挖掘效果上的提升。

第四章：介绍了基于RBF神经网络和用户画像的混合推荐算法。针对数据稀疏性这个问题，提出使用RBF神经网络优化评分矩阵，预测空缺的评分项。然后提出了一种将RBF神经网络和用户画像模型相结合的新的混合推荐算法，并针对新提出的算法进行了试验验证，论证了该算法的提升效果。

第五章：基于用户画像的移动广告推荐系统的设计与实现。结合第三章与第四章对用户画像建模方法和推荐算法的改进，依据现有的项目设计并实现了基于用户画像的移动广告推荐系统。从需求分析、架构设计、结构设计、数据库设计与实现、系统测试等方面给出了详细的阐述。

第六章：总结和展望。全面总结了本文所做的工作，展望本文提出的算法与系统在将来需要改进的方向。

# 第二章 相关理论与关键技术

## 2.1 引言

本文将运用机器学习、数据挖掘等前沿技术，并且这些技术已相对成熟，并且还在不断往前发展。

## 2.2 机器学习和数据挖掘

机器学习（Machine Learning, ML）是当前人工智能的研究热点，融汇了概率论、统计学、逼近论、凸分析、计算方法等多门学科的内容，主要研究计算机如何模拟或实现人类的学习行为，以获取新的知识或技能，重新组织已有的知识结构使之不断改善自身的性能。目前，机器学习已经应用在了很多领域，比如淘宝网的商品推荐、文字识别、语音识别，人脸识别、医学分析等。机器学习的应用使得其应用领域智能化，简单化，更为有效的为用户提供服务。

数据挖掘（Data Mining, DM），它是数据库知识发现（Knowledge-Discovery in Databases，KDD)中的一个步骤。数据挖掘一般是指从大量的数据中通过算法搜索隐藏于其中信息的过程。数据挖掘通常与计算机科学有关，并通过统计、在线分析处理、情报检索、机器学习、专家系统（依靠过去的经验法则）和模式识别等诸多方法来实现上述目标。

## 2.3 用户画像

## 2.4 推荐算法

## SSH开发相关技术

# 第三章 相关数据挖掘

## 3.1 引言

## 3.2 数据挖掘方法

## 3.3 数据挖掘结果分析

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、吉屋网等国内排名前五的房产网站的数据爬取，数据量达20多TB。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。

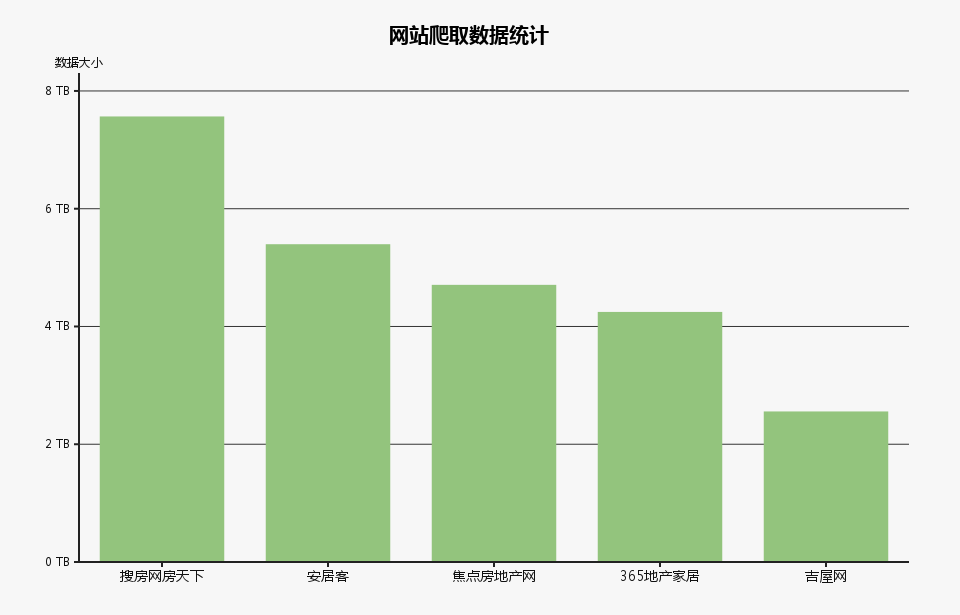


图1　 爬取数据统计图

通过分析得到国内大部分城市的房价分布，根据房价分布可以对房屋所属小区的档次做一个大致判断，然后再采集小区周围的一些地理特征。以武汉是2015-2016年的房价分布为例，可以看出汉口平均房价高于其他几个地区的房价，分析其地理特征可以发现，汉口临江，并且还是武汉的金融、商业、贸易中心，地处繁华地带。由此我们可以初步判断，该片地区的人们的消费能力应该是普遍高于平均消费水平的。

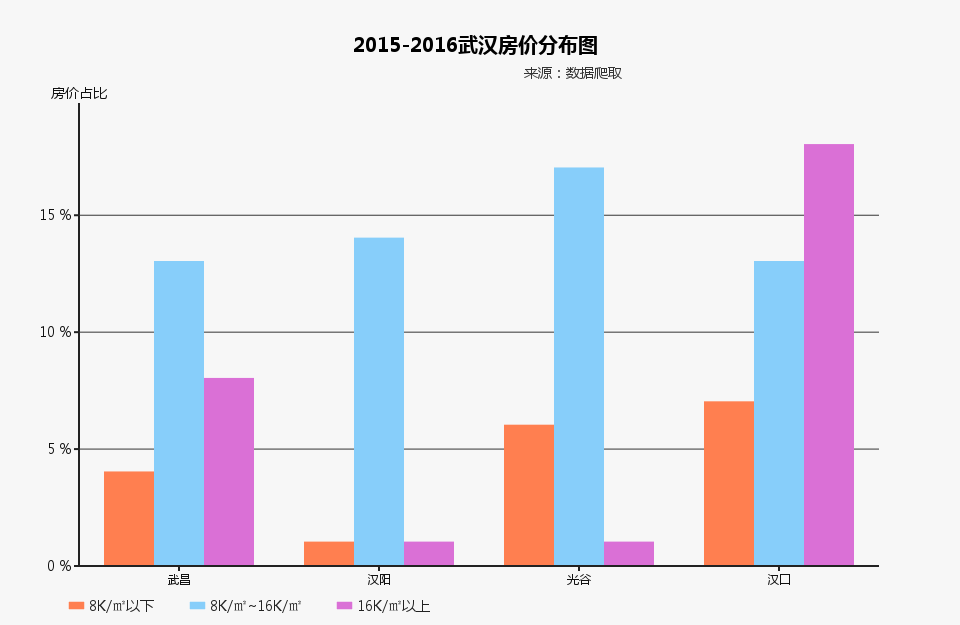


图2　 2015-2016武汉房价分布图

## 3.4 本章小结

# 第四章 用户画像模型的建立与优化

## 4.1 引言

## 4.2 用户画像流程

## 4.3 用户画像建模

## 4.4 实验结果分析

## 4.5 本章小结

# 第五章 基于用户画像的推荐算法

## 5.1 引言

## 5.2 推荐算法

## 5.3 协同过滤

## 5.4 基于用户画像的协同过滤推荐算法

## 5.5 本章小结

# 第六章 广告精准投放系统的设计与实现

## 6.1 引言

## 6.2 需求分析和系统设计

### 6.2.1 市场需求分析

对于城市中的广告位招商，仍采用线下的传统方式进行：广告投放商自主选择**→**广告制作商进行广告印制**→**由物业进行定点投放广告。这种方式有着诸多缺点。一方面，广告投放商获得的可投放的广告位的选择可能是比较狭窄，从而导致广告投放商的广告投放位置可能与广告内容不相适应，致使广告的作用降低，达不到最初设想的广告效果。另一方面，对于广告位的拥用者物业公司而言，他们有可能有空闲的广告位却得不到及时的安置，这无疑造成了极大的浪费。针对这样充满缺点的线下广告位交易手段，本项目计划将这一广告位交易的程序移植到互联网中，将交易中各方的信息进行精准对接，在此基础上，实现对广告投放商的个性化推送服务，更进一步去满足广告投放商多样化需求，提高广告的投放效率，节省双方的资源。

### 6.2.2 用户需求分析

项目针对用户主要有两类人：广告投放商和广告位所有方（多数情况为物业公司，下文以物业公司代替）。

广告投放商参与到广告位交易的最终目的是为了宣传自己的产品，因此，广告投放商的主要需求就是实现广告作用的最大化。为此，本项目将开展如下工作：

a)收集尽可能多的广告位信息并将它们提供给广告投放商，并可以对这些广告位信息进行筛选，以更好地满足广告投放商的要求；

b)根据广告投放商的以往购买行为作为依据，运用机器学习的知识，来实现为广告投放商智能的推送一些个性化消息。

至于物业公司，其本意是为了增加自己的利益，充分利用空闲的广告位，以便于充分利用手中资源获取最大化的利益。为了保障物业公司的利益，将根据物业公司所在楼盘的价格，推算出该片区受众的消费能力，然后把相应物业公司的广告位推送给与之适应广告投放商。

## 6.3 系统的设计

前端展示技术：Sass层叠样式表语言, Vue数据驱动界面技术, jQuery多浏览器兼容技术,gulp自动任务运行器。后台技术框架: Spring框架、SpringMVC框架, RESTful接口风格。数据库：mongodb分布式文件存储数据库，redis内存数据库。

本阶段的主要工作是将产品需求转化为设计需求，指导后续的编码工作。设计需求要求阐述了产品需求的详细设计方案，包括页面布局、数据结构、算法以及易用性、安全性、可扩展性、健壮性和性能等诸多方面的设计思路。在正式编码前，针对需求写出相应的《软件功能表》来指导后续的编码工作。这样做有两大好处：一是在编码之前就充分预见到将来可能遇到的问题，可以尽早规避风险；二是为开发工作搭好框架，降低因开发人员的差异导致开发过程的不确定性，避免出现“一千个人心中有一千个对需求的理解”。

针对我们的项目，我们项目的核心在于智能化地为客户提供建议，提供个性化的服务，运用机器学习的有关知识，系统根据客户之前已完成的交易，对现有的商品进行基于商品的协同过滤，或者系统将会筛选出具有类似购买行为的客户，然后进行基于客户的协同过滤，为具有类似购买行为的客户，进行相互的商品推荐。

项目组目前已经利用爬虫程序对各大房产交易网站的交易数据进行了收集，这些数据将会被用来进行受众的成分提取，让我们了解受众的收入情况，从而使得之后的广告投放变得更加精准。

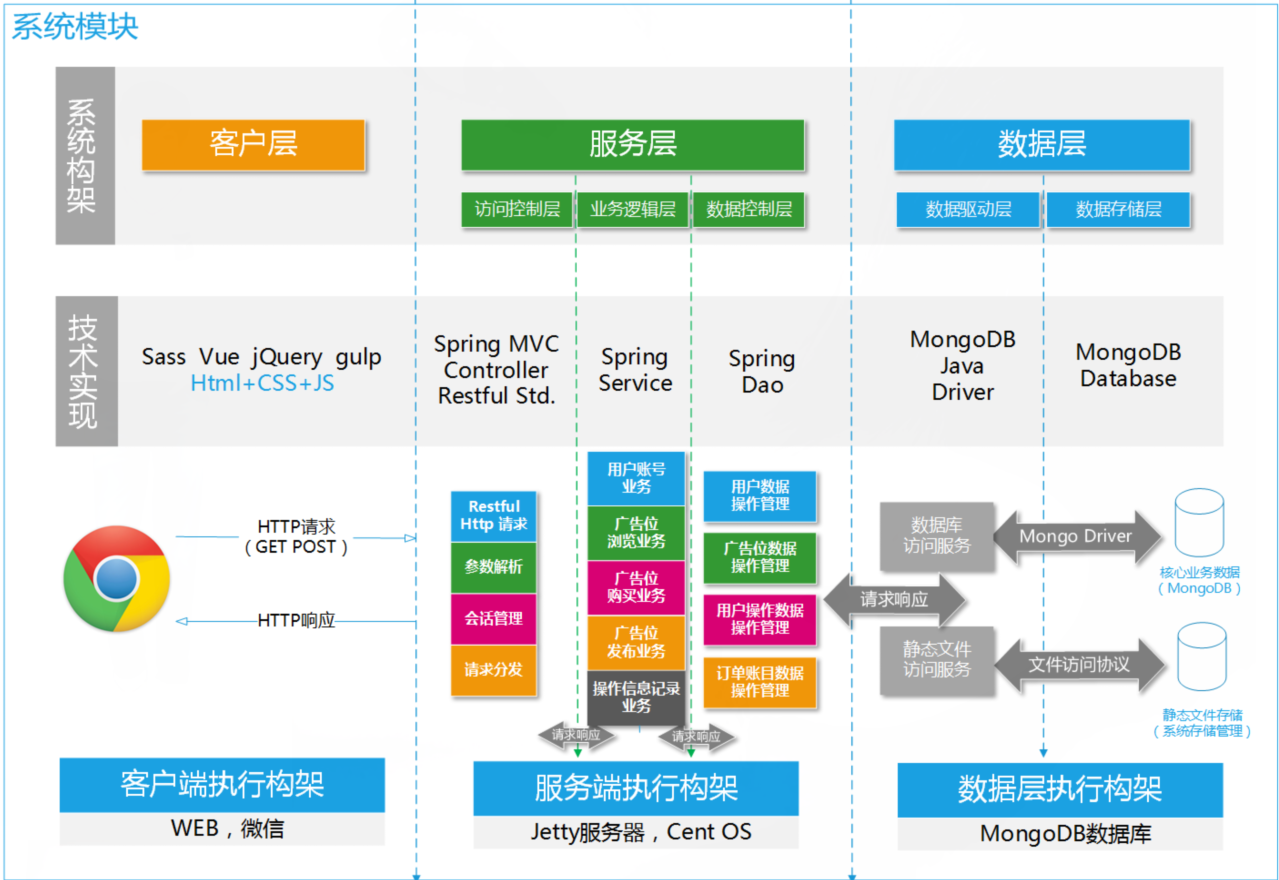


图9 系统模块图

系统模块如图9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，使用Sass和gulp来辅助开发。

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Spirng 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。数据存储层使用MongoDb来管理保存数据。

## 6.3.1 开发环境介绍

## 6.3.2 开发技术介绍

## 6.3.3 广告精准推荐模块的实现

## 6.4 系统展示

# 第七章 总结与展望

二、立项背景

2.1 市场需求分析

对于城市中的广告位招商，仍采用线下的传统方式进行：广告投放商自主选择**→**广告制作商进行广告印制**→**由物业进行定点投放广告。这种方式有着诸多缺点。一方面，广告投放商获得的可投放的广告位的选择可能是比较狭窄，从而导致广告投放商的广告投放位置可能与广告内容不相适应，致使广告的作用降低，达不到最初设想的广告效果。另一方面，对于广告位的拥用者物业公司而言，他们有可能有空闲的广告位却得不到及时的安置，这无疑造成了极大的浪费。针对这样充满缺点的线下广告位交易手段，本项目计划将这一广告位交易的程序移植到互联网中，将交易中各方的信息进行精准对接，在此基础上，实现对广告投放商的个性化推送服务，更进一步去满足广告投放商多样化需求，提高广告的投放效率，节省双方的资源。

2.2 用户需求分析

项目针对用户主要有两类人：广告投放商和广告位所有方（多数情况为物业公司，下文以物业公司代替）。

广告投放商参与到广告位交易的最终目的是为了宣传自己的产品，因此，广告投放商的主要需求就是实现广告作用的最大化。为此，本项目将开展如下工作：

a)收集尽可能多的广告位信息并将它们提供给广告投放商，并可以对这些广告位信息进行筛选，以更好地满足广告投放商的要求；

b)根据广告投放商的以往购买行为作为依据，运用机器学习的知识，来实现为广告投放商智能的推送一些个性化消息。

至于物业公司，其本意是为了增加自己的利益，充分利用空闲的广告位，以便于充分利用手中资源获取最大化的利益。为了保障物业公司的利益，将根据物业公司所在楼盘的价格，推算出该片区受众的消费能力，然后把相应物业公司的广告位推送给与之适应广告投放商。

三、项目方案

3.1 项目设计方案

3.1.1 拟采用的相关技术

项目将运用机器学习、数据挖掘等前沿技术，并且这些技术已相对成熟，并且还在不断往前发展。

（1）数据分析技术

机器学习（Machine Learning, ML）是当前人工智能的研究热点，融汇了概率论、统计学、逼近论、凸分析、计算方法等多门学科的内容，主要研究计算机如何模拟或实现人类的学习行为，以获取新的知识或技能，重新组织已有的知识结构使之不断改善自身的性能。目前，机器学习已经应用在了很多领域，比如淘宝网的商品推荐、文字识别、语音识别，人脸识别、医学分析等。机器学习的应用使得其应用领域智能化，简单化，更为有效的为用户提供服务。

数据挖掘（Data Mining, DM），它是数据库知识发现（Knowledge-Discovery in Databases，KDD)中的一个步骤。数据挖掘一般是指从大量的数据中通过算法搜索隐藏于其中信息的过程。数据挖掘通常与计算机科学有关，并通过统计、在线分析处理、情报检索、机器学习、专家系统（依靠过去的经验法则）和模式识别等诸多方法来实现上述目标。

（2）系统构建技术

前端展示技术：Sass层叠样式表语言, Vue数据驱动界面技术, jQuery多浏览器兼容技术,gulp自动任务运行器

后台技术框架: Spring框架、SpringMVC框架, RESTful接口风格

数据库：mongodb分布式文件存储数据库，redis内存数据库

3.1.2 系统体系架构

本项目的体系架构如图3所示。

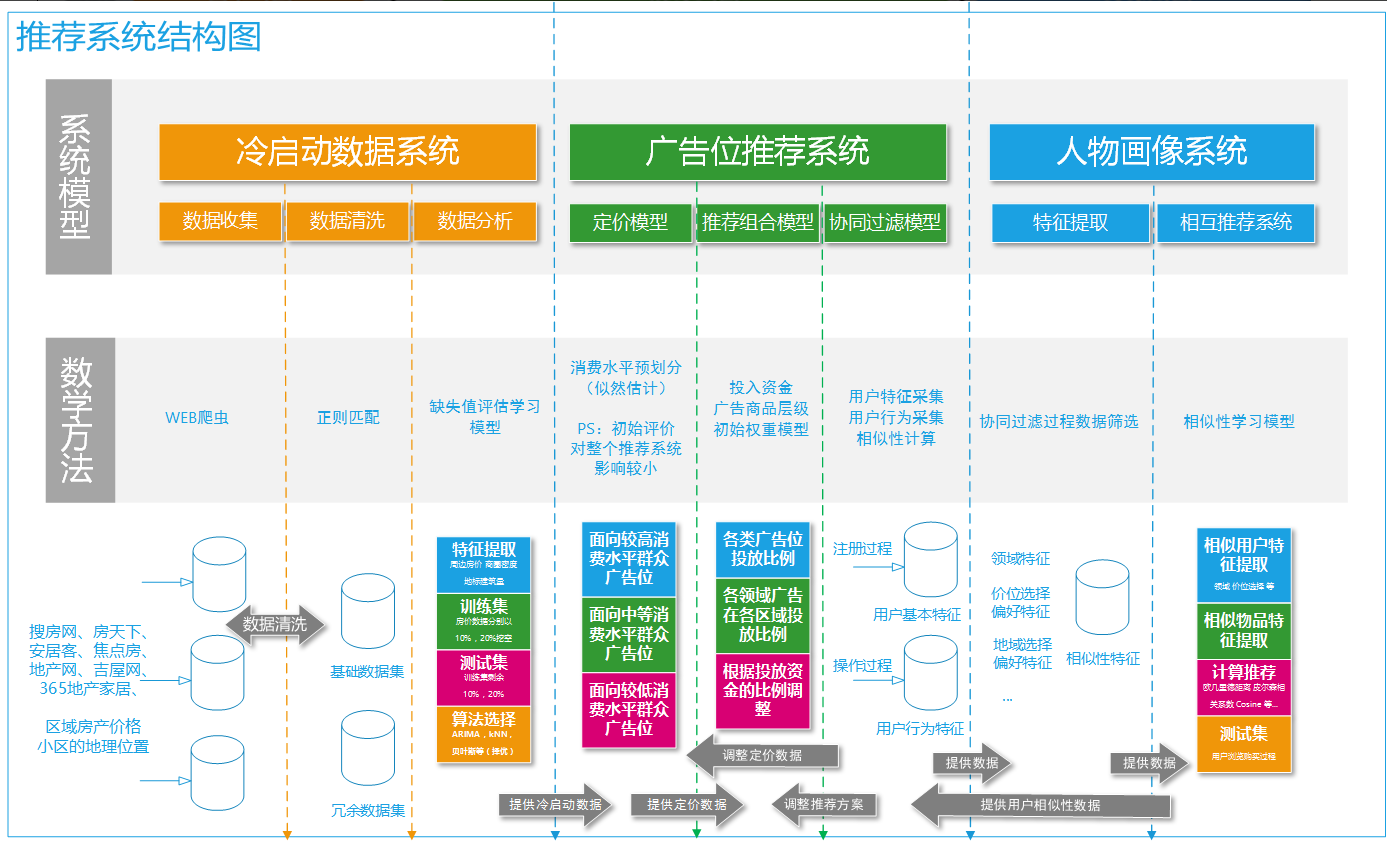


图3 推荐系统结构图

1. 冷启动数据系统
   1. 数据采集：web爬虫爬取搜房网、房天下等网站的房地产价格数据
   2. 数据清洗：对过时数据，失效数据通过正则匹配等方式进行清洗
   3. 数据分析：
      * 目的：通过与周边相关性评估缺失数据的小区房产价格
      * 特征提取：房价、商圈密集度、地域
      * 训练集：已知房价以10%挖空作为训练集
      * 测试集：挖空数据做测试集
      * 算法选择：KNN，贝叶斯等（择优）
2. 广告位推荐系统
   1. 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
   2. 推荐组合模型（确定投放方案）
      * 评价标准：广告商资本量，投放产品领域，投放产品消费档次，及广告位价格
      * 目的：确定广告商在各个消费群体的投放比例和地域分布
   3. 协同过滤模型
      * 用户特征采集：注册过程
      * 用户行为采集：操作过程（浏览，购买等）
      * 相似性分析：向人物画像系统提供数据取得相似性信息，为推荐组合模型提供精准投放方案
3. 人物画像系统
   1. 特征提取
      * 特征筛选：主要考虑投放商领域，偏好价位，偏好地域等
   2. 相互推荐
      * 计算用户相似性
      * 计算广告位相似性
      * 计算推荐
        1. 为相似投放商提供相似方案
        2. 将相似广告位推荐给购买过该类广告位的投放商

3.1.3核心问题解决方案与算法

1. 待解决核心问题：

1.制定有针对性的投放方案，面对不同领域的投放商制定不同的投放组合。

2.对广告位根据层次进行定价，实现物业效益最大化。

1. 已达成的前置条件：

与武汉市广告生产投放中介的合作

1. 待建立的基本投放模型：

领域A的广告投放商，在时间区间B内（主要考虑季节，天气等对特定需求广告的影响），在地域C范围中，对不同档次D的小区（群体）的投放比例

1. 可直接提取的特征
2. 地域C的层次特征，划分依据：来自搜房网房天下、安居客、焦点房地产网、365地产家居、吉屋网的该区域房产价格以及小区的地理位置等。
3. 广告投放商的工作领域A，划分依据：由广告投放商自行选择。
4. 可根据服务计算理论建立模型提取的特征
5. 小区档次D（群体层次）特征模型：根据所属的地域C特征以及通过LBS平台得到的周边商圈发展程度。

e.g : 地域C靠近江边，区域内多为200万以上豪华套房且该周边商圈数量庞大，通过这些特征，可以预估该小区消费水平较高。

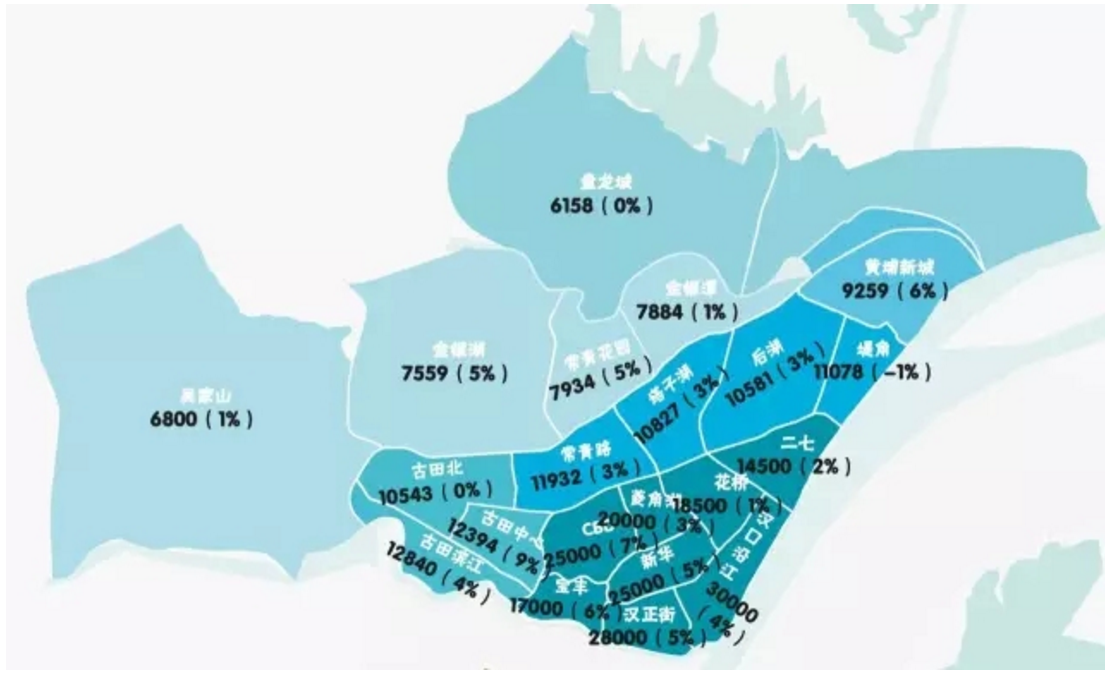


图4 武汉汉口房价分布情况图

图4是2016年5月武汉汉口商品住宅的价格分布情况，从图中可以看出，汉口房价的分布情况大致是离江距离越近价格越高，距离中心商业区越近房价越高，选择在这些区域买房的人的消费能力肯定也是较高的，所以我们可以把这些区域的小区划分到较高档次。

1. 时间区间B对领域A广告的影响：根据该领域广告商在该时间段内的投放情况：

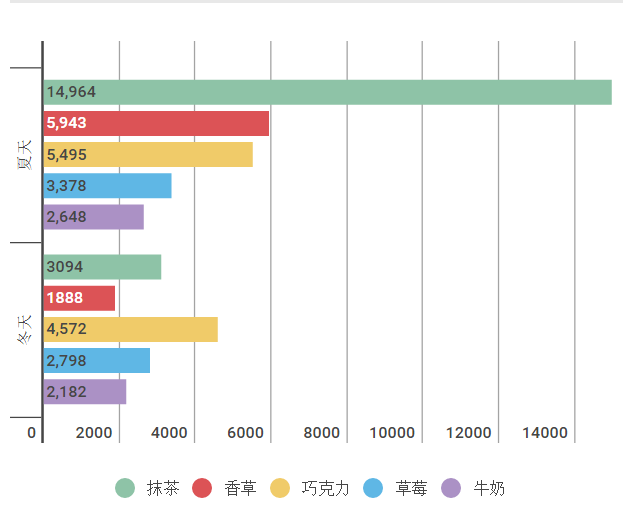


图5 冰淇淋广告投放商的时间-销量图

从图5可以看出，冰淇淋广告投放商在冬季的投放需求明显减少.

1. 通过机器学习算法初步训练得到的理想模型

准备数据：广告商领域A，时间区间B内A的投放情况，地域范围C内房价商圈情况。

训练及测试算法：

1. 聚类分析：根据地域情况（道路，河流），以及商圈密集程度，通过聚类分析算法，区分不同的地域范围C，并通过采样分析进行算法调整。

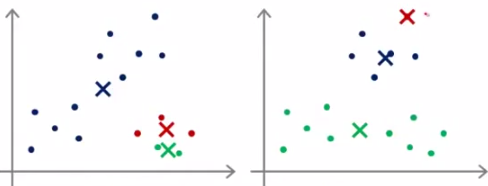


图6 算法优化回归图

1. 服务计算：根据已有的区域C内房价数据以及地域特征，评估各个小区的群体层次D，对广告位进行定价，并通过采样分析进行算法调整。
2. 建立简要bp神经网络（图7），对已有特征数据（领域A，时间B内A类广告的投放状况，群体层次D）进行训练，初步获得分类模型，获得时间段B，层次D对领域A的服务需求程度。

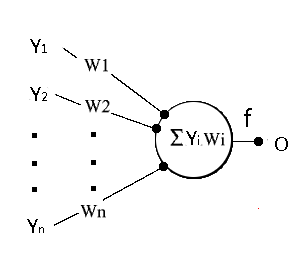


图7 bp神经网络节点图

1. 构建投放策略：领域A的投放商，在时间B内，以当前有限的资金情况对小区D1、D2、D3…的广告投放比例
2. 通过用户协同过滤得到的优化模型

由于特征无法考虑周全，在原有推荐模型的基础上，对领域A广告投放商的广告位选择操作进行统计分析，优化原有模型，调整各个参数在实际生产中的比例，达到精准投放的效果。

（8）图谱数据分析

建立图8所示的广告关系模型，进一步挖掘影响广告投放的因素。

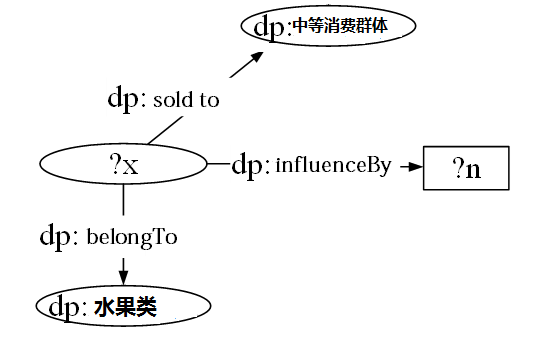


图8 广告关系模型图谱

3.1.2可能出现的障碍及解决方案

（1）数据过于稀疏：对于部分偏远区域，由于数据量过少，无法进行有效的聚类分析。将建立统一标准，根据用户协同过滤作用做进一步调整。

（2）无法预估模式性事件和突发事件对广告影响：节假日等模式性的事件，和台风等突发性事件。对模式性事件进行录入，对突发性的事件进行分析并保存，在出现相同模式前预警。

（3）模型偏离较远：爬虫频繁访问等情况导致出现参数大幅度偏离。根据现有服务框架，模型参数将会实时保存，随时回滚。

3.2项目实施方案

此阶段处于整个项目实施工作的最前期，由成立项目组、前期调研、编制总体项目计划、启动会四个阶段组成。其中：

指导老师：在合同签定后指定项目经理，成立项目组，授权项目组织完成项目目标。

项目组：进行前期项目调研，与用户共同成立项目实施组织，编制《总体项目计划》，召开项目启动会。

此阶段主任务：

1. 成立项目组：

指导老师通过对组内成员的进行一定了解，任命项目经理，指定项目目标，由项目经理和指导老师一起指定项目组成员及成员任务，由项目经理编写《项目任务书》并要求队员签署。

2. 编制《项目总体计划》：

《项目总体计划》是一个文件或文件的集合，随着项目信息不断丰富和变化，会被不断变更，主要介绍项目目标、主要项目阶段、里程碑、可交付成果。通常包括以下几方面内容：

项目描述，沟通管理计划，确定项目干系人对信息和沟通的需要：即什么人何时需要什么信息以及通过什么方式将信息提供给他们。

质量管理计划，确定适合于项目的质量标准和如何满足其要求。如果有必要，可以包括上述每一个计划，详细程度根据每个具体项目的要求而定。

未解决事宜和未定的决策。

3.2.1初代产品开发阶段

此阶段包含需求、设计、编码、测试四个步骤，四个步骤构成一个开发周期。最开始的第一个周期所得到的项目成果对项目后期的进展影响很大，所以需要进行严密的设计，充分考虑项目投入应用后可能出现的问题。所以该阶段在很大程度上决定了我们项目的成功与否。

3.2.1.1 需求

本阶段的主要工作是需求制定与评审。该阶段的工作分三步走：

收集原始需求**→**制定产品需求**→**产品需求评审。

首先通过各种渠道收集原始需求，由于原始需求多半是概念性的、模糊的，不能直接用来指导开发工作，所以需要进行归类、筛选，整合为产品需求。基本原则是，结合当前开发产品的特性，争取以最小的改动以及最大的可扩展性来制定产品需求。降低风险，同时提高灵活性。经验告诉我们，在需求没有考虑透彻的情况下，不要贸然开始设计并实现，可能导致大量返工，费时费力。产品需求制定好后，制定组员编写《需求文档》，交由指导老师进行评审。

针对我们的项目而言我们需要做的是，分析物业方、广告投放方、广告制造方的三方关系，然后再对单独的各方实现各自的功能需求。在初步对我们的项目进行分析后，我们暂时确定的功能具有下面8个方面：

（1）受众群体特征分析（数据挖掘）

在进行特征分析时，采取LBS（基于位置服务），基于对象成分划分等方式，针对异源多构的数据模型进行多种数据分类，采取有效的算法和优先级的调控，给出最合适的受众群体的选择。

LBS（基于位置服务）特征分析。LBS（基于位置服务）是指通过电信移动运营商的无线电通讯网络或外部定位方式，获取移动终端用户的位置信息，在GIS平台的支持下，为用户提供相应服务的一种增值业务。LBS的关键地方在于PaaS和BaaS的服务模式的有效结合，提供给用户更加吸引人的服务。在项目准时时期，我们经过长期的市场调研，发现了地域特征与用户群体之间存在的关系，根据不同特征的用户聚集，进行用户的特征分析。

基于对象成分划分的特征分析。社会阶层是具有相同或类似社会地位的社会成员组成的相对持久的群体。研究发现，同属一个社会群体中的用户往往具有一些共同特征，比如经济消费能力，价值观念等等。我们采用的基于对象成分划分的特征分析，一定程度上是建立在不同群体所对应的不同消费能力与消费对象的基础上，进行的深入分析。

（2）广告商特征分析（数据挖掘）

进行特征分析时，采取LBS（基于位置服务），商业类型等方式，针对医院多购的数据模型进行多种数据分类，采取有效的算法和优先级的调控，给出最合适的广告商的选择。

LBS（基于位置服务）特征分析。对于广告投放商来说，其所在的位置信息与投放广告形成的影响能力有着很大的关系。我们基于位置服务，对于广告投放商进行特征分析，根据投放商种类，建立足够的数据模型，可能优先选择投放地为投放商所在位置的周边地区，增强投放的有效性。

商业类型特征分析。商业类型对于投放地有着很大的影响，比如，文教类产品在学术区的投放效果就会明显好于居住区。基于这种商业作用的直接模式，建立复杂的模型，完成商业类型的特征分析。

（3）分散/聚集模型的建立，根据商家要求进行量化投资和分散投资策略的合理评估

量化投资是一种操作方法或操作理念，与其他各种“非量化”的方法并列。两话也可以采取择时/趋势跟踪/超跌/强弱对冲等等投资模型。区别仅在于，量化投资会使用量化的行情和走势来进行买卖点决策，而不是传统的图形式行情。分散化是投资策略中的一条准则，也就是“不要把鸡蛋都放在一个篮子里”。分散化是指把资金分散投资，与企业多元化有相同的本质。通过建立分散/聚集的资金分配模型，可以根据商家的要求进行分散投资策略的评估或者直接给出投资指导，实现投资效益的最大化。

（4）基于用户的协同过滤和基于商品的协同过滤

基于用户的协同过滤是指，经过大量的统计数据后，对于具有相同特征或者相同优先级的商家，根据其做出的选择，提供相应的广告投放位置信息推荐。基于用户的协同过滤的优点：

① 能够过滤难以进行机器自动基于内容分析的信息。

② 能够基于一些复杂的难以表达的概念（用户特征）进行过滤。

③ 推荐的新颖性。

基于商品的协同过滤是指，根据上面讲到的用户的特征分析，过滤掉不符合广告商需求的受众群体，给出合适的投放位置信息推荐。基于商品的协同过滤的优点：

① 首先给广告商提供一个基于数据模型的广告投放位置信息推荐。

②广告商在进行最终广告位的选择时，也对于该模型进行了优化。

③ 经过大量的数据统计后，可以得到用户数据与广告商数据的信息库。

（5）利用数据库完成用户画像与广告商画像

在经过大量的模型优化与数字统计后，会积累大量的用户特征分析信息与广告商的特征分析信息。通过对于这些信息的提取分类，组合成所有用户与广告商的数据信息，使得这两种模型立体化。所用到的数据可以用于后期进行数据分析和现象的研究，最终建构完整的知识理论体系，在适当的理论推动下，推进该行业更好的发展。

（6）多平台的实现模式

①WEB端的软件设计

WEB作为主流平台，开发难度较小，应用面更广。本项目趋向于设计完整的中间商交易平台，整合基于机器学习与大数据的精准广告投放系统，完成项目的初步应用。

②去app的发展模式

当前，截止到2015年第一季度，微信已经覆盖中国 90% 以上的智能手机，月活跃用户达到 5.49 亿，用户覆盖 200 多个国家、超过 20 种语言。此外，各品牌的微信公众账号总数已经超过 800 万个，移动应用对接数量超过 85000 个，微信支付用户则达到了 4 亿左右。如此庞大的用户规模，为该项目的进行提供了有利的基础。本项目的宗旨在于去app的发展模式，摒弃过去庞大复杂的应用模式，选用新一代的微信应用。新一代的微信应用相对传统的app具有如下优点：

* 易开发：小程序比APP的开发难度降低至少一半，还能跨平台，且接近原生的体验，大幅减少了应用开发的人力，降低创业门槛。
* 易更新：小程序基于Web，可以即时更新（如不考虑微信的审核），有效加速产品的迭代。
* 易互通：小程序之间通过链接就可以相互调起、传递数据。
* 易搜索：小程序中的数据方便搜索引擎索引，可以打破APP信息孤岛。
* 易使用：应用无需安装卸载，即点即用。
* 易传播：传播链接或二维码即可。
* 易推广：利用微信强大的社交流量，获客成本低且更精准。
* 易变现：微信支付的基础设施已经搭好，用户的支付习惯已经养成，微信的广告生态也日益成熟

（7）对于广告实际安装投放的奖励式结算方式

采用微信应用的方式，对于每完成一个地方的广告安装，由安装者拍照上传，即可在线完成工资的清算，摆脱过去复杂的资金分流方式，省去了繁琐的中间环节。

3.2.1.2 设计

本阶段的主要工作是将产品需求转化为设计需求，指导后续的编码工作。设计需求要求阐述了产品需求的详细设计方案，包括页面布局、数据结构、算法以及易用性、安全性、可扩展性、健壮性和性能等诸多方面的设计思路。在正式编码前，针对需求写出相应的《软件功能表》来指导后续的编码工作。这样做有两大好处：一是在编码之前就充分预见到将来可能遇到的问题，可以尽早规避风险；二是为开发工作搭好框架，降低因开发人员的差异导致开发过程的不确定性，避免出现“一千个人心中有一千个对需求的理解”。

针对我们的项目，我们项目的核心在于智能化地为客户提供建议，提供个性化的服务，运用机器学习的有关知识，系统根据客户之前已完成的交易，对现有的商品进行基于商品的协同过滤，或者系统将会筛选出具有类似购买行为的客户，然后进行基于客户的协同过滤，为具有类似购买行为的客户，进行相互的商品推荐。

项目组目前已经利用爬虫程序对各大房产交易网站的交易数据进行了收集，这些数据将会被用来进行受众的成分提取，让我们了解受众的收入情况，从而使得之后的广告投放变得更加精准。

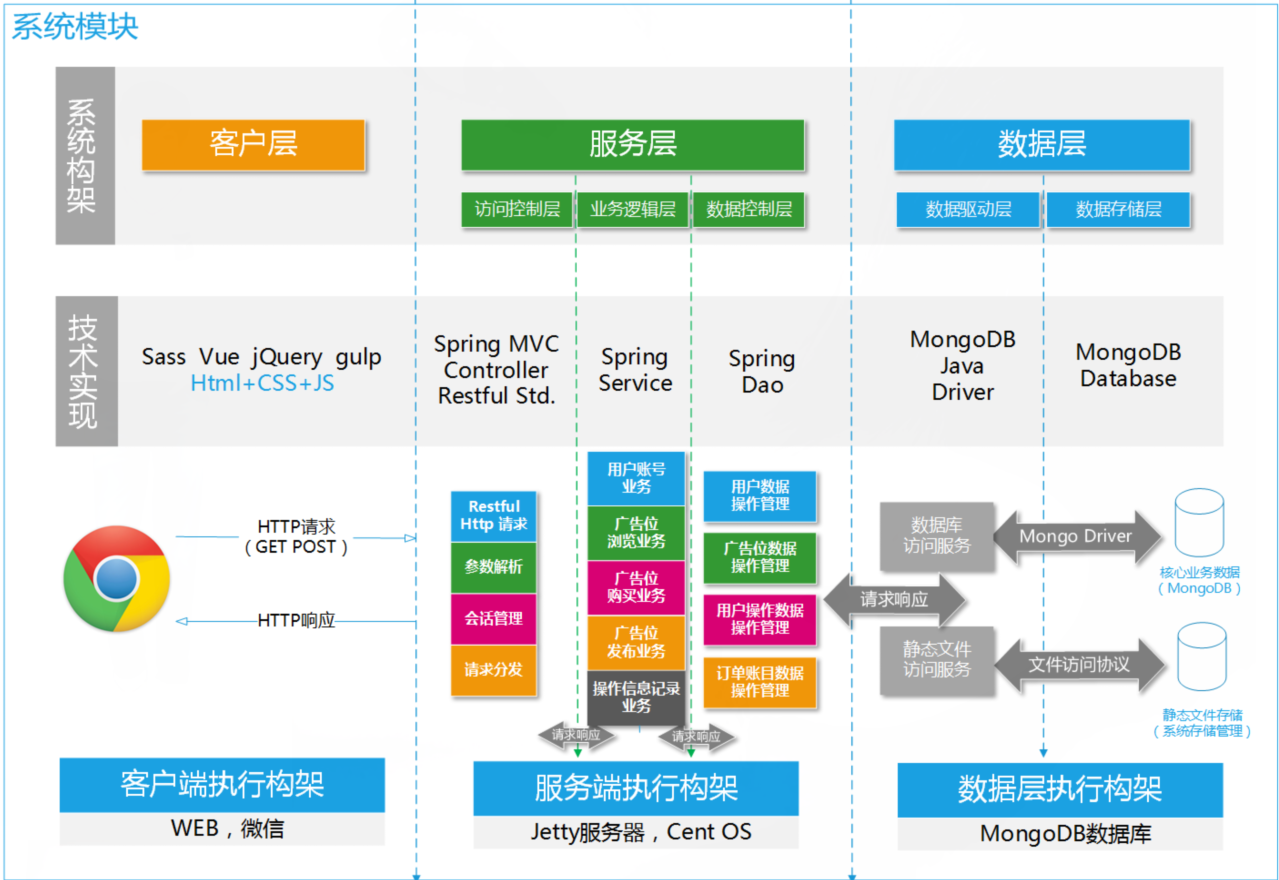


图9 系统模块图

系统模块如图9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，使用Sass和gulp来辅助开发。

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Spirng 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。数据存储层使用MongoDb来管理保存数据。

3.2.1.3 编码

本阶段的主要工作是严格按照《软件功能表》来完成编码，并组织进行代码评审，填写《软件功能确认表》。代码评审的重点应该是对程序结构的审查，发现深层次的软件错误，而不要停留在表面。尽量保证，每一个编码-测试迭代里，都要完成相对独立、可测试性强的功能点。

3.2.1.4 测试

本阶段的主要工作是，在每一个测试构建中，尽可能多地覆盖需求点，并根据轻重缓急合理安排测试优先级，尽可能将影响较大的缺陷提前暴露出来。测试优先级的安排应遵循以下原则：

a、先测试经过变更的部分，然后测试没有变更的部分

b、先测试程序的核心功能，然后测试一般功能

c、先测试逻辑性的功能，然后测试业务性的功能

d、先测试常规情况，然后测试异常情况

e、先测试功能，然后测试性能

通过观察测试人员使用情况，和相关人员进行沟通，记录并编写《测试及试运行总结》。

3.2.2迭代开发阶段

重复需求、设计、编码、测试的开发周期，不断迭代开发。每一个迭代都可以形成一个可交付的小版本。每个迭代周期内，对于编码和测试也可进行多次迭代。通过快速发布测试构建的方式，验证开发完成的新功能，再通过测试发现问题来驱动开发人员对软件进行修改完善，循环往复。即：根据开发情况有针对性地组织测试，根据测试结果反作用于开发人员去完善软件质量。以这种小步快跑的方式，经若干测试构建后，软件质量可以在较短时间内达到稳定状态。

3.2.3验收阶段

此阶段是对项目总体的完成情况进行验收。验收分阶段进行，在每一项目阶段结束时，用户对这一阶段的可交付成果进行验收，在测试及试运行结束后，对系统进行总体验收。