分类号 _____ 密 级 _____

U D C _____ 编 号 10486

武漢大學

硕 士 专 业 学 位 论 文

基于大数据分析的广告精准投放研究

研 究 生 姓 名：王焰

学 号：2015212163317

指导老师姓名、职称：郑宏 教授

专 业 名 称：软件工程

研 究 方 向：软件工程

二〇一八年五月

A Study on Accurate Advertising of Advertising Based on Big Data Analysis

By

Yan Wang

Supervised by

Prof. Hong Zheng

Wuhan University

Wuhan，430079 P.R.China

May，2018

郑 重 声 明

本人的学位论文是在导师指导下独立撰写并完成的，学位论文没有剽窃、抄袭、造假等违反学术道德、学术规范和侵权行为，本人愿意承担由此而产生的法律后果和法律责任，特此郑重声明。

学位论文作者（签名）：

年　　月　　日

摘　要

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。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。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。

在上述客观条件下，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。

关键词：大数据，协同过滤，数据挖掘，广告投放，推荐系统

Abstract

With the increasing popularity of mobile devices, mobile advertising occupies more and more market share. Compared to the traditional Internet, mobile media has its own characteristics such as mobile, fragmented and individualized, which requires the development of mobile advertising in the direction of precision and individuation. Only by changing the extensive mode of the extensive advertising of traditional advertising and personalized advertising for different interests of different users, can the advertising be converted into consumer behavior, so that both advertisers and advertisers can get good commercial returns. However, most of the existing personalized advertising recommendations are based on content recommendation, first extracting the key words of the user's current page, and then putting in the ads that match it without taking into account the interest of the user itself.

This project uses the latest machine learning technology, through the collection and mining of large data such as real estate, region and surrounding business circle, create association rules and cluster analysis of different levels of audience to establish the advertising model. In addition, through the collaborative filtering and combination analysis of advertising location selection by different advertising providers, the advertisers can accurately deliver to the users. This project will build specific knowledge base based on the acquired big data, and further tap the deep connection between user groups and advertisements.

Key words: Big Data, Collaborative Filtering, Data Mining, Advertising, Recommender System

目 录

第一章 绪论

本章主要介绍文章的选题背景和意义，通过背景内容的介绍和相关研究现状的研究，引出广告行业的问题，指出推荐系统在广告中的重要作用，进而提出基于数据分析的广告投放精准推荐系统，最后指出本文的研究内容和目标。

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），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，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。

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]。在当前的大背景下，移动终端在互联网设备中占据主导地位，互联网广告投放的研究以移动端为主，而移动端广告具备以下的几个特性。

移动性：广告是附着在移动设备上的，因此广告具备移动性，用户的地理位置不断变换，因此根据用户的地点，可以考虑投放不同类型的广告。

互动性：移动设备上的应用大部分是可交互的，终端用户和设置之间有相应的交互界面。例如移动终端上的 HTML5 游戏，整个应用不断地与用户交互，用户产生的行为持续比较长的时间。

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。

扩散性:移动广告具有很强的扩散性，互联网信息传播模型驱动广告信息的传播。利如用户浏览到感兴趣的广告，可以通过社交媒体如微信、微博等进行转发，从而实现信息的扩散[1][5]。

可追踪：每个移动设备与用户是绑定的，所以设备与用户之间有严格的映射关系，因此广告背后用户的信息是可以精确追踪的。

上述特性决定了传统粗放型的广告投放方式不再适用于个性化和精细化要求比较高的移动广告投放，因此研究基于大数据和用户行为的精准广告投放具有实际应用价值。

1.2 相关技术的发展和研究现状

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

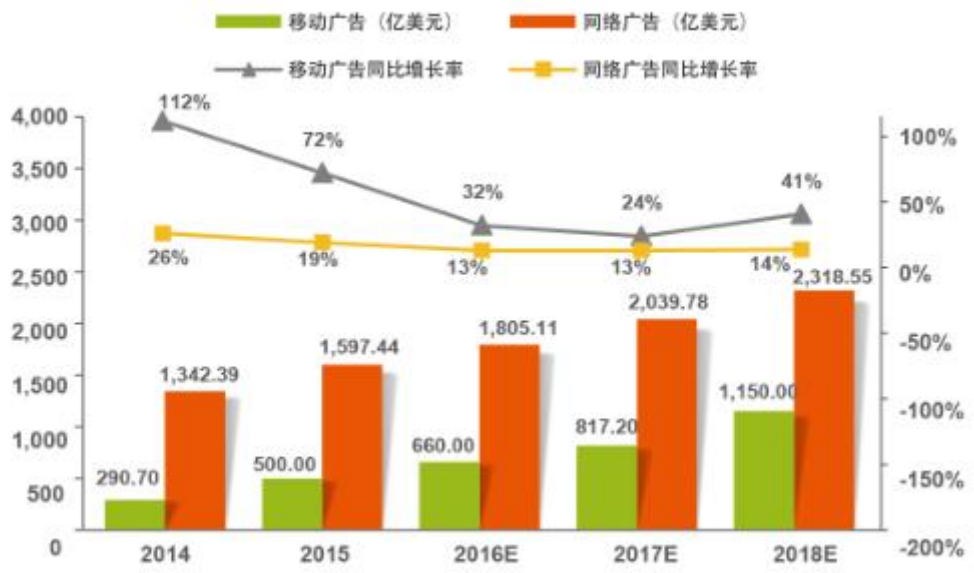

图 1 2013-2018 年全球移动广告和网络广告市场规模及增长率

近年来，随着大数据和云计算的成熟，机器学习和人工智能的发展，使得原有的广告行业发生了天翻地覆的变化，其中互联网广告行业由于其本身的特性受到了更多的关注，在前景发展迅猛。2015 年，互联网广告延续了之前的高增长态势，但增幅正在放缓，远远低于上年度 51.7%的增长，增幅回落到 35.3%。随后的几年，广告的增长趋势趋于平稳，整体的增长规模也呈现出稳健的增长势态。

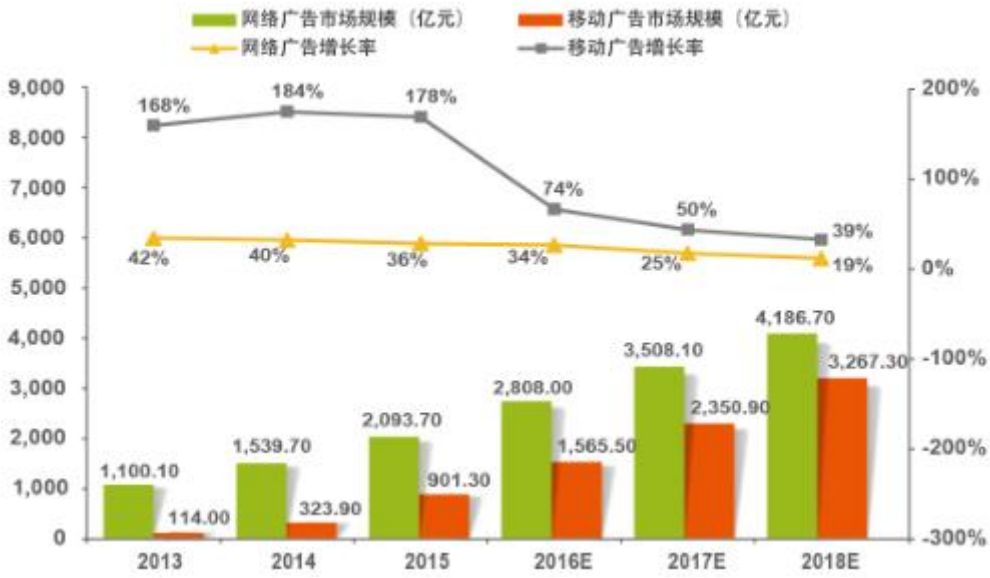

图 2 2013-2018 年中国网络广告和移动广告市场规模情况

据不完全统计，截止到 2017 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。移动端应用为移动端互联网广告的投放提供了前期的基础。

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。

用户行为分析是利用数据挖掘技术获取某一商品的用户行为数据，对这些用户行为数据进行统计分析，从内部找到用户行为的规则，将所发现的规律应用到网络营销策略中，并应用于连续更新网络营销策略。

从精确的广告投放和用户行为分析的定义和特点，可以看出，精确的广告投放本质上是通过数据挖掘技术，挖掘网络用户行为的数据，找到用户的爱好，并针对这些用户的喜好进行广告匹配，最终实现精准广告。

准确的广告投放作为网络广告最流行的一种形式，满足了用户的个性化需求，节约了公司成本，规范了用户与企业的关系。自创建以来，它一直是国内外研究的热点。然而，由于互联网在西方国家的发展相对较早，无论是理论研究还是企业应用，国外均处于领先位置。

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，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。

1.3 研究内容

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。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。

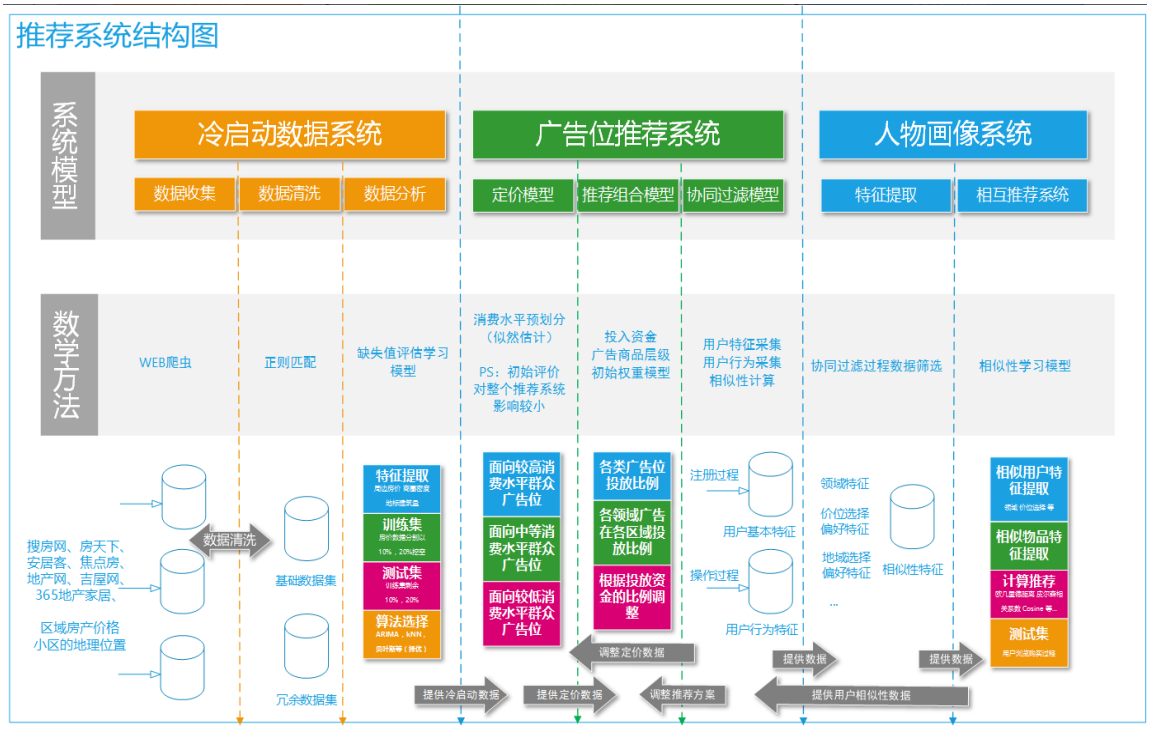

图 3　推荐系统结构图

本推荐系统共有三个模块构成，即冷启动数据系统模块、广告位推荐系统模块、人物画像系统模块。

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

1.4 论文结构

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，先介绍了本课题的研究背景和研究意义，然后对本课题相关技术的发展和研究现状进行了介绍，最后对本文的研究内容进行了详细的介绍。

第二章：介绍相关的理论和关键技术。先介绍了机器学习和数据挖掘的概念，然后介绍了用户行为，最后介绍了推荐系统中的系统过滤算法。

第三章：介绍了基于爬虫等数据挖掘的方法，爬取了信息房产和用户的相关信息，并对数据和信息进行本章首先介绍了系统的体系结构，给出了系统的总体框架。然后从各个模块出发介绍了各个模块的功能，模块算法的大致的步骤。。

第四章：主要讨论了数据采集，数据清洗，数据量化，然后通过算法计算出地域的广告投放价值和人群购买力价值。构建广告投放精准推荐系统相关算法，算法的核心是协同过滤算法，同时还考虑到推荐系统中会遇到冷启动问题，本文采取利用历史数据、利用用户偏好和设置问答的方式来解决广告推荐中的冷启动问题。

第五章：主要介绍了广告投放系统的实现和实验评估。在第三章与第四章对用户画像建模方法和推荐算法的基础上，对系统进行了实现，包括需求分析、系统设计和系统功能的实现，然后设计了相关实验，最后通过实验对系统性能进行了分析和评估。

第六章：总结和展望。全面总结了本文所做的工作，分析了本文工作存在的不足并对未来的工作方向进行了展望。

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。同时对用户行为进行了分析，用户行为分析包括广告商分析、广告类型分析和广告受众分析。协同过滤是推荐算法中最常见最有效的，分为基于用户的协同过滤和基于物品的协同过滤，主要通过相似性分析，来推荐相似的邻居。

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，以概率论、逼近论、凸优化和数值计算方法为理论，把计算机模仿和学习人类的行为作为研究目标，并在整个学习的过程中不断改进自身的性能。主目前，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。

大数据正深刻影响着人们的生产方式、生活习惯、思维模式和研究方法。大数据不仅是学界和业界的前沿课题，而且已上升为国家基础性战略资源。大数据的独特之处，除了规模巨大、类型多样、增长迅速等特性，最重要的是这些特性所导致的"全息"意义上的数据关联性，这种关联性将是实现未来商业模式、生产生活方式、管理流程等颠覆性变化的驱动力。数据关联性也是导致常规的数据保护与隐私保护方式失效的根本原因之一。例如，关联性挖掘分析使得仅通过匿名技术不能很好地保护用户隐私。但是，如果施加过强的数据保护策略，必将割裂这些数据的关联性，从而形成一个个数据孤岛并导致大数据服务的不可用。

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。数据挖掘真的是从大量的数据当中获取规律和知识的方法。数据挖掘是一本计算机应用科学，目前的应用领域包括在线分析

处理、情报检索、和专家系统等，数据挖掘已经成为解决当前计算机领域中信息检索问题中最常见的理论。

2.3 用户行为

2.3.1 用户行为的概念

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，如何收集用户的偏好信息成为系统推荐效果最基础的决定因素。用户有很多方式向系统提供自己的偏好信息，而且不同的应用也可能大不相同。

2.3.2 用户行为的分类

现在互联网用户的主要接入媒介为：个人电脑、智能手机、平板电脑等终端。而用户行为一般可以分为以下几类：

(1)眼动行为。对眼动行为的研究在国外还是比较常见的，外国学者对此研究要求比较高，并且也取得了较高的研究成果。在中国也有一些学者开始研究眼动行为。对用户眼动行为的研究，可以了解到用户对哪些东西感兴趣，哪些界面布局合适或者不合适。另外，通过改进用户的操作界面可以提高用户的体验。

(2)鼠标点击与移动行为。自从出现鼠标后，用户在网上最多的操作行为就是鼠标行为，所以对用户行为的分析离不开分析鼠标行为。鼠标的行为主要是鼠标点击和鼠标移动。当前国内外有很多成熟的系统，可以记录和分析鼠标移动和点击行为。除此之外，第三方公司也可以为中小公司提供鼠标点击和移动行为信息。

(3)操作键盘的行为。因为鼠标不能大量输入信息，键盘通常可以输入大量信息，键盘输入行为是数据分析的内容之一，不能忽略。

(4)其他设备的触摸和点击等行为。这些设备通常可以代替鼠标和键盘的工作。现在新的触摸和点击技术可以产生很复杂的用户行为，我们也要对这些行为进行分析和研究。

表 2-1 用户行为和用户偏好

用户行为	类型	特征	作用
评分	显式	整数量化的偏好，可能的取值是 [0, n]；n 一般取值为 5 或者是 10	通过用户对物品的评分，可以精确的得到用户的偏好
投票	显式	布尔量化的偏好，取值是 0 或 1	通过用户对物品的投票，可以较精确的得到用户的偏好
转发	显式	布尔量化的偏好，取值是 0 或 1	通过用户对物品的投票，可以精确的得到用户的偏好。如果是站内，同时可以推理得到被转发人的偏好（不精确）
保存书签	显示	布尔量化的偏好，取值是 0 或 1	通过用户对物品的投票，可以精确的得到用户的偏好。
标记标签(Tag)	显示	一些单词，需要对单词进行分析，得到偏好	通过分析用户的标签，可以得到用户对项目的理解，同时可以分析出用户的情感：喜欢还是讨厌
评论	显示	一段文字，需要进行文本分析，得到偏好	通过分析用户的评论，可以得到用户的情感：喜欢还是讨厌
点击流（查看）	隐式	一组用户的点击，用户对物品感兴趣，需要进行分析，得到偏好	用户的点击一定程度上反映了用户的注意力，所以它也可以从一定程度上反映用户的喜好。
页面停留时间	隐式	一组时间信息，噪音大，需要进行去噪，分析，得到偏好	用户的页面停留时间一定程度上反映了用户的注意力和喜好，但噪音偏大，不好利用。
购买	隐式	布尔量化的偏好，取值是 0 或 1	用户的购买是很明确的说明这个项目它感兴趣。

以上的用户行为的划分是比较粗糙的，在实际的推荐引擎设计中，需要根据本身的特性自定义用户的行为，从而更精确地表达用户的行为方式。

在实际的应用当中，不同的行为需要进行组合，用户行为的组合模式有两种方法：

（1）将不同的行为分组：例如浏览和下单等，然后基于不同的行为，计算不同的用户或物品相似度。类似于淘宝或京东给出的“购买了该物品的人还购买了 …”，“查看了宝贝的人还查看了 …”

（2）根据不同行为反映用户喜好的程度将它们进行加权，主要分为显式的用户喜好和隐式的用户喜好，显式的用户喜好更能直接的表达用户的兴趣，因此占的比重较高，同时相应的权值也就比较大。

收集的用户数据需要进行相应的处理，其中最核心的工作就是：减噪和归一化。

（1）减噪：在用户行为数据中可能存在噪声数据，因此需要根据一定的信息过滤方法把噪声过滤掉，这样可以使收集到的用户行为数据更加精确。

（2）归一化：为了使得加权求和得到的用户行为数据更加精确，需要进行归一化处理。最基本的归一化处理，就是将数据的值除以数据取值的上界，数据被映射到了[0,1]范围中。

在上述预处理的基础上，根据用户行为数据，可以构建用户偏好的矩阵，同时商品也可以更见相应的商品矩阵，这两个矩阵中的值是对用户行为的抽象，根据这两个矩阵中可以进行相应进行推荐。

进行的预处理后，根据不同应用的行为分析方法，可以选择分组或者加权处理，之后我们可以得到一个用户偏好的二维矩阵，一维是用户列表，另一维是物品列表，值是用户对物品的偏好，一般是 [0,1] 或者 [-1, 1] 的浮点数值。

2.4 协同过滤

2.4.1 推荐系统

推荐系统(Recommendation System, RS)，简单来说就是根据用户的日常行为，自动预测用户的喜好，为用户提供更多完善的服务。推荐系统在理论描述或

实际应用中，已经非常广泛并且成为一套体系。推荐系统的定义：用U(User)表示用户集合，用I（Item）表示可以推荐给用户的对象的集合，推荐系统就是为用户集合找出项目集合中可以满足用户需求的项目，可以用映射函数来表示：

$$\int : U \times I \rightarrow R \qquad (2.1)$$

函数表示项目I对用户U的推荐满意度，R是推荐范围内的非实数队列项目一般用评分表示接受程度，评分是指单个用户在多大的程度上喜欢某个项目，越喜欢则评分越高。然而一般来说，接受程度不仅指评分，也可以是任意函数，如利润函数等。用户集U中的每个用户都可以用一些特征进行对象化，例如性别、年龄、婚姻状态、民族、爱好、收入等，在系统中最直观省事的方式可以用用户ID来表示。类似地，I中的每一个项目也可W由一系列特征进行对象化处理，这取决于项目是什么。比如，如果项目是广告，那就可W表示成广宵名称、广告类型、广告主、广告时间、广告佣金等。从上述定义和模型可看出，用户建模、推荐对象建模和推荐算法三个方面是推荐系统的关键技术。

协同过滤推荐算法是诞生最早，并且较为著名的推荐算法。主要的功能是预测和推荐。算法通过对用户历史行为数据的挖掘发现用户的偏好，基于不同的偏好对用户进行群组划分并推荐品味相似的商品。协同过滤推荐算法分为两类，分别是基于用户的协同过滤算法(User-based Collaborative Filtering)，和基于物品的协同过滤算法(Item-based Collaborative Filtering)。简单的说就是：人以类聚，物以群分。下面我们将分别说明这两类推荐算法的原理和实现方法。

推荐系统使用了一系列不同的技术，主要可以分为以下两类：

基于内容(Content-Based，CB)的推荐。主要依据的是推荐项的性质。

基于协同过滤(Collaborative Filtering， CF)的推荐。主要依据的是用户或者项之间的相似性。

在协同过滤方法中，我们很显然的会发现，基于协同过滤的推荐系统用可以分为两类：基于项(Item-Based，IB)的推荐系统。主要依据的是项与项之间的相似性。基于用户(User-Based，UB)的推荐系统。主要依据的是用户与用户之间的相似性。

当已经对用户行为进行分析得到用户喜好后，我们可以根据用户喜好计算相似用户和物品，然后基于相似用户或者物品进行推荐，这就是最典型的 CF 的两

个分支：基于用户的 CF 和基于物品的 CF。这两种方法都需要计算相似度，下面我们先看看最基本的几种计算相似度的方法。

2.4.2 相似度的计算

相似度的计算有多种方法，下面本文将详细介绍几种常见的计算相似度的理论。

（1）欧几里德距离（Euclidean Distance）

最初用于计算欧几里德空间中两个点的距离，假设 x，y 是 n 维空间的两个点，它们之间的欧几里德距离是：

$$d(x,y)=\sqrt{(\sum(x_i-y_i)^2)} \tag{2.2}$$

可以看出，当 n=2 时，欧几里德距离就是平面上两个点的距离。

当用欧几里德距离表示相似度，一般采用以下公式进行转换：距离越小，相似度越大

$$sim(x,y)=\frac{1}{1+d(x,y)} \tag{2.3}$$

（2）皮尔逊相关系数（Pearson Correlation Coefficient）

皮尔逊相关系数一般用于计算两个定距变量间联系的紧密程度，它的取值在 [-1，+1] 之间。

$$p(x,y)=\frac{\sum x_iy_i-n\overline{xy}}{(n-1)s_xs_y}=\frac{n\sum x_iy_i-\sum x_i\sum y_i}{\sqrt{n\sum x_i^2-(\sum x_i)^2}\sqrt{n\sum y_i^2-(\sum y_i)^2}} \tag{2.4}$$

sx，sy 是 x 和 y 的样品标准偏差。

（3）Cosine 相似度（Cosine Similarity）

Cosine 相似度被广泛应用于计算文档数据的相似度：

$$T(x,y)=\frac{x\bullet y}{\|x\|^2\times\|y\|^2}=\frac{\sum x_i y_i}{\sqrt{\sum x_i^2}\sqrt{\sum y_i^2}} \tag{2.5}$$

（4）Tanimoto 系数（Tanimoto Coefficient）

Tanimoto 系数也称为 Jaccard 系数，是 Cosine 相似度的扩展，也多用于计算文档数据的相似度：

$$T(x,y)=\frac{x\bullet y}{\|x\|^2+\|y\|^2-x\bullet y}=\frac{\sum x_i y_i}{\sqrt{\sum x_i^2}+\sqrt{\sum y_i^2}-\sum x_i y_i} \tag{2.6}$$

2.4.3 相似邻居的计算

介绍完相似度的计算方法，下面我们看看如何根据相似度找到用户 - 物品的邻居，常用的挑选邻居的原则可以分为两类：图 1 给出了二维平面空间上点集的示意图。

（1）固定数量的邻居：K-neighborhoods 或者 Fix-size neighborhoods 不论邻居的“远近”，只取最近的 K 个。

（2）基于相似度门槛的邻居：Threshold-based Neighborhoods

在这种方法中，邻居的数量是不确定的。在计算的过程中首先设置一个门阀，只有超过这个阈值才被考虑为相关的邻居节点。这种方式的优点是计算出来的相似度不会有太大的误差。如图 1 中的 B，从点 1 出发，计算相似度在 K 内的邻居，得到点 2，点 3，点 4 和点 7，这种方法计算出的邻居的相似度程度比前一种优化，尤其是对孤立点的处理。

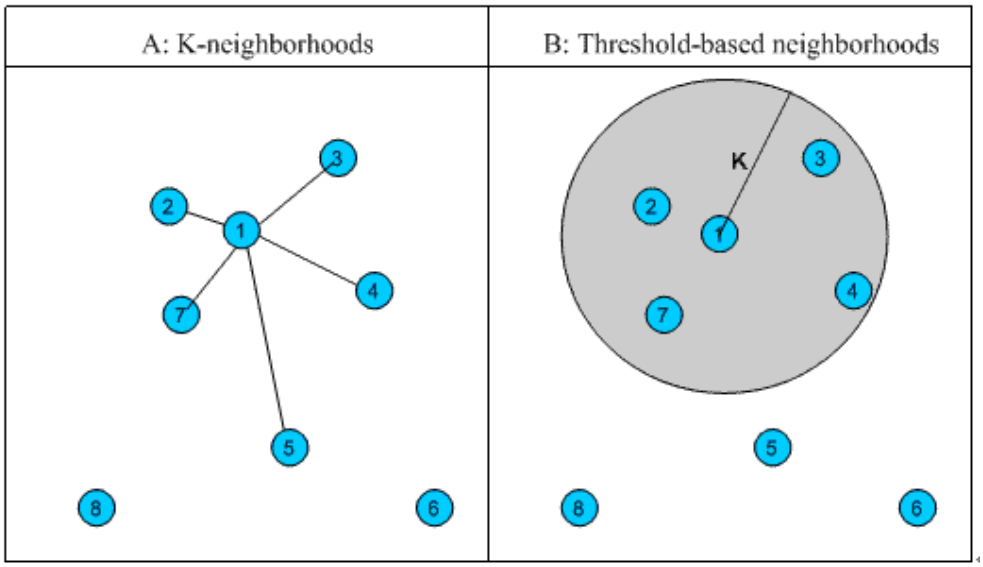

图 2-1 相似邻居计算示意图

基于用户的协同过滤算法是根据用户的历史行为数据对用户兴趣相似度计算，然后推荐用户兴趣度较高的商品，这种方式能够发现兴趣相似度比较高的用户，然后推荐两个用户相关的产品。利用用户 A 和用户 B 都购买了 x、y 和 z 三个产品，并且都有相同的评价，那么就可以认为用户 A 和用户 B 是相似度比较高的用户，此后把用户 A 购买过的商品 W 推荐给 B 是合理的行为。

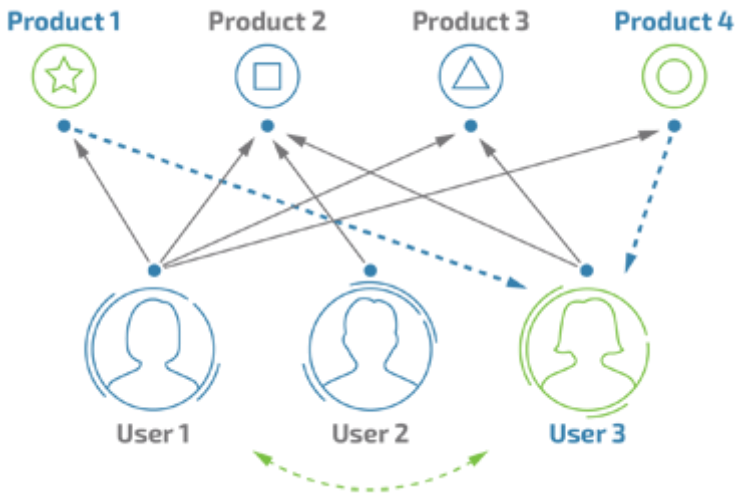

图 2-2 协同过滤过程

2.5 本章小结

本章系统的介绍了本文中要用到的相关理论和关键技术，其中包括机器学习

和数据挖掘，通过数据挖掘技术挖掘文章需要的信息和数据，主要包括用户数据和广告数据，运用机器学习中方法进行数据分析，通过对用户行为的分析，提取用户的特征和广告的特性，并利用这些特征建立模型。主要利用协同过滤的算法和方法来进行广告的推荐，为广告商提供最好的投放方案。

第三章 地缘特征数据挖掘和分析

3.1 引言

根据第二章相关的技术和理论的分析可以知道，精准广告的投放更加需要对用户行为和广告投放目标进行相关的数据收集和分析。利用数据挖掘技术，从网上爬取相应的地缘特征数据，主要根据城市房价信息和商圈信息，同时对用户的行为信息进行抓取。为后续相关系统的建立做数据准备。

3.2 数据爬取

网络爬虫，又称网页蜘蛛和网络机器人，是按照一定的规则，自动抓取网页的计算机程序。爬虫通常从某一个起始页开始抓取网页，读取网页的内容，解析出其中的链接，再通过这些链接寻找新的页面，这样一直循环，直到满足系统的终止条件而停止抓取[7]。

3.2.1 网络爬虫

网络爬虫分为通用网络爬虫和聚焦爬虫。

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，解页面，处理析网页的内容，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。

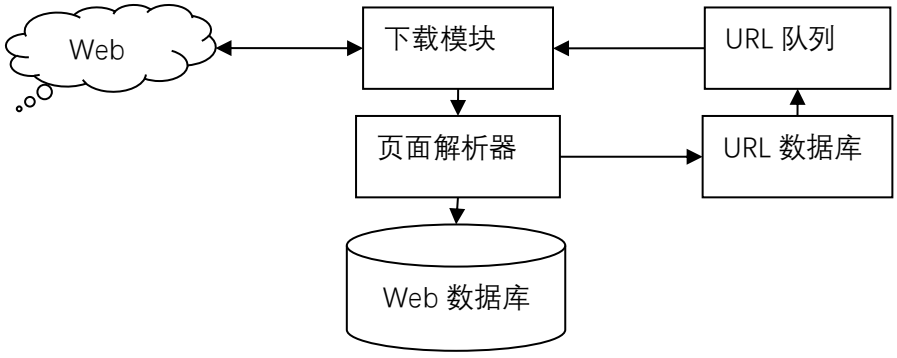

图 3-1 通用网络爬虫的结构图

而聚焦爬虫是在通用爬虫的基础上添加一些主题定制模块，因此它除了调度器，解析器和资源库之外，一般还要有搜索策略和网页及 URL 的主题相关度评价模块。本文设计的聚焦爬虫结构图见图 3-1.

3.2.2 爬虫流程

通用网络爬虫常常是简单的下载页面内容，追求的是对于网络的搜全率，要求数据资源很庞大全面。在过滤方面有些常常是简单的为 URL 添加关键字进行页面过滤，如文献[16]。而聚焦爬虫针对的是某个领域的数据，相反，追求的是精细而专注的数据，因此它就要求有一定的网页和链接的分析过滤方法。

下面就从通用爬虫和聚焦爬虫的工作流程来进行比较[1]，如图 2-2 所示。

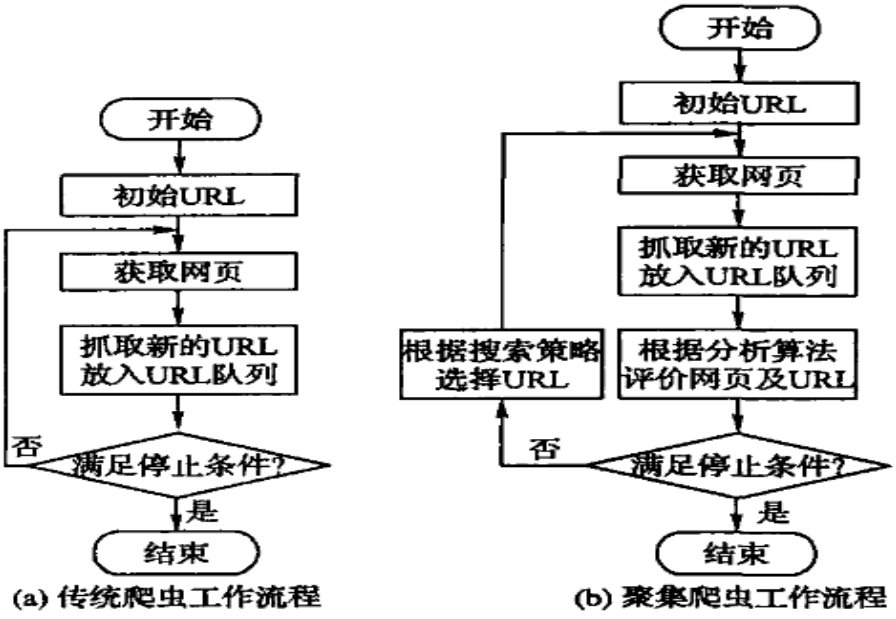

图 3-2 传统爬虫和聚焦爬虫的工作流程对比

从图 3-2 可以看出，传统爬虫和聚焦爬虫的区别，聚焦爬虫需要对网页进行过滤，所以对它所抓取的 URL 都要根据一定的分析算法进行评价，以过滤符合主题的网页。

由于通用模块是由控制器，解析器和资源库三部分组成，而聚焦爬虫是在通用爬虫的基础上增加主题相关模块来实现的，所以聚焦爬虫比通用爬虫要多出一些模块用于定制主题[14]。

目前，聚焦爬虫主要有三种代表性的体系结构：基于分类器的聚焦爬虫[5]，

基于数据抽取器的聚焦爬虫和基于用户学习的聚焦爬虫[15]。本文所研究的聚焦爬虫正是一个侧重于数据抽取的聚焦爬虫，因此采用了基于数据抽取器的聚焦爬虫体系结构。系统框架图如图 3-3 所示。

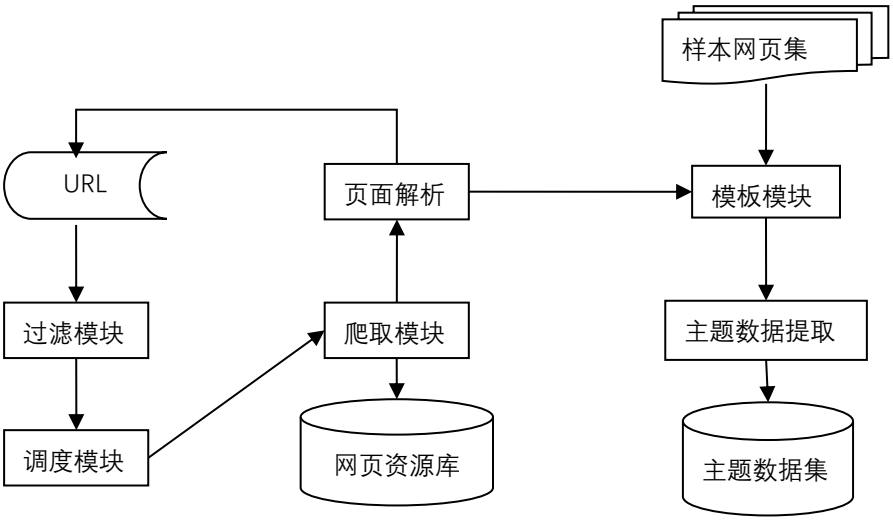

图 3-3 系统框架图

3.2.3 模块细解

由系统框架图 3-4 可知，此聚焦爬虫的重点是对主题数据的抓取，因此主题定制和模版匹配就是聚焦爬虫的思想体现所在。而模版匹配必须得有一个好的页面解析算法解析网页，才能准确高效的进行匹配，获得所需的数据。本节将详细介绍各个模块的组成和功能等。

3.2.3.1 搜索过滤模块

搜索过滤模块主要是负责按照一定的搜索策略进行 URL 的搜索，然后对搜索到的 URL 进行过滤，获得有用的链接。搜索到的每个 URL 都要经过过滤，再决定是否抓取。

过滤主要是负责 URL 的过滤，它的任务就是对 URL 进行主题相关性评价，使得那些和主题相关的链接保留下来，从而舍弃无用的链接。要实现 URL 的过滤功能，依靠用户的过滤设置和对目标网页的链接结构描述。

过滤设置包括所要下载的文件的类型、大小等，文件类型主要有图形文件（比如 jpg、gif、png 等）、文本文件（如 html、txt、asp 等）、可运行文件（如 exe、

com 等)、压缩文件（如 zip、rar 等)、音频文件（mp3、wav、wma 等)、视频文件（avi、mov、wmv 等)。过滤还包括 URL 的深度等参数。

对目标网页的链接结构描述，是指指定所要查找的相关主题网页的链接所具有的共同结构，通常也就是他们都有什么样的链接前缀。这样的策略是基于如下的思想：相同主题的网页通常都有着很大的相似性，包括页面结构及连接结构，他们通常都是从同一个祖先链接（比如指向某个目录页面的链接）通过一层或多层链接而来的[16]。

经过了过滤而得到的 URL 都放在一个优先队列中，该优先队列按页面入度值进行排序。然后调度模块会给队列中的每个 URL 分配线程，进行抓取。

3.2.3.2 调度和爬取模块

调度模块是爬虫的控制器，负责分配线程。系统同时开启的线程数由用户指定，系统根据此设定开启指定数目的线程。对由过滤模块得到的 URL 候选队列中的每一个 URL，调度模块分配给它一个线程，启动爬虫抓取模块进行抓取。

调度模块因为是对线程的管理，所以它要实现数据之间的同步和共享数据的安全问题。比如 URL 候选队列，每个线程都要访问它来获得一个候选 URL 进行抓取和解析等，如果一个 URL 被好几个线程同时获得，会导致对页面的重复提取。所以共享数据的共享和同步很重要，保护好这些共享数据，才能避免重复工作，数据冲突和其它因同步引起的问题。多线程是比较复杂的编程问题，调度模块为了实现多线程，应提供一个线程池， 对每个请求分配一个线程，对共享资源以互斥的方式操作，并避免死锁等问题[17]。

爬取模块根据调度模块的请求，采集一个网页，并更新网页资源库，同时抽取该网页的链接给过滤模块[18]。它的功能比较简单，它主要负责下载网页，网页下载后，要传给解析模块进行解析，然后根据用户的指定（存储与否），来决定是否存储在硬盘里。在网页下载存储过程中，需要解决的一个问题即是编码问题，不同的网页有不同的编码，所以对下载的网页获取它的编码，以便在后续的页面解析中产生乱码问题等。

3.2.3.3 HTML 解析模块

解析模块是爬虫核心模块，它需要对爬取模块爬取的内容进行解析。解析模块主要有词法分析器和语法分析器两部分组成。如图 3-5 所示：

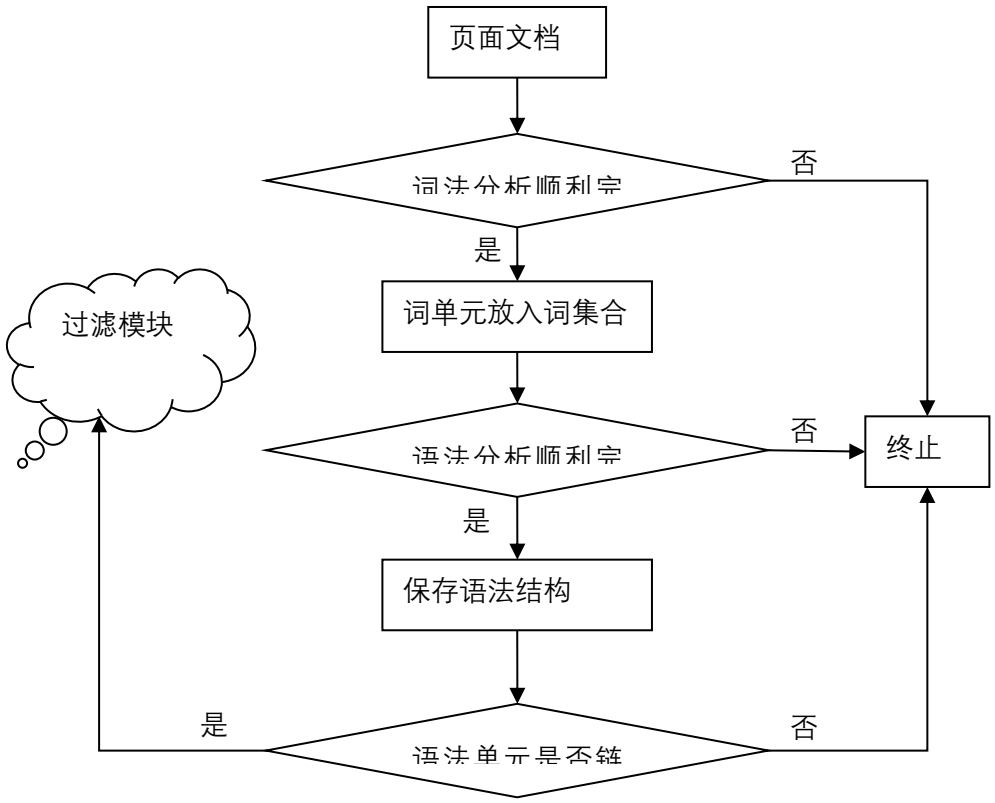

图 3-4 网页解析流程图

如图 3-4 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。

然后语法分析器对词法单元进行语法分析，解析出其语法结构并保存起来。对于每个语法单元，语法分析器会调用一系列监听器来监听语法单元的内容。链接监听器会判断该语法单元是否是一个链接。如果是链接，就将其传给过滤模块以判断其是否和主题相关。对于其它的语法单元，语法分析器会解析出它的语法成分，以供以后的主题建模及模板匹配使用。

由于现在的浏览器都有一定的容错性，所以很多网页的 HTML 源码虽能正确显示，但其中很多都有着一定的词法或语法错误。所以，如果要获得这些网页资源并抽取信息，自己实现的 HTML 分析器就必须也得具备一定的容错能力，能在一定的错误范围内解析出源码相应正确的语法结构[19]。关于本文的 HTML 分析器

如何实现容错功能。

3.3 数据挖掘方法

（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。

3.4 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。

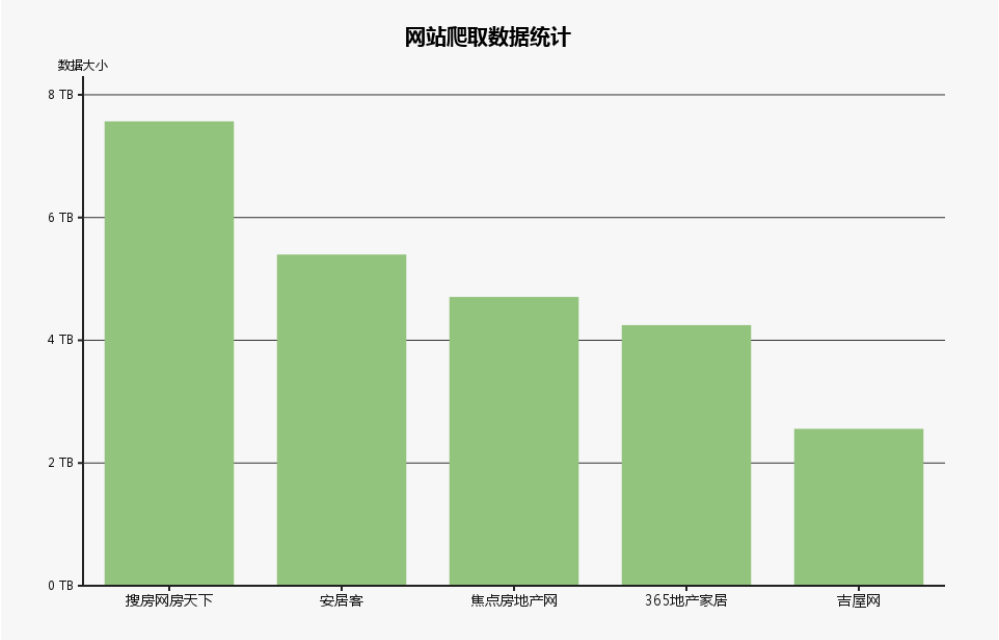

图 3-5　爬取数据统计图

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。

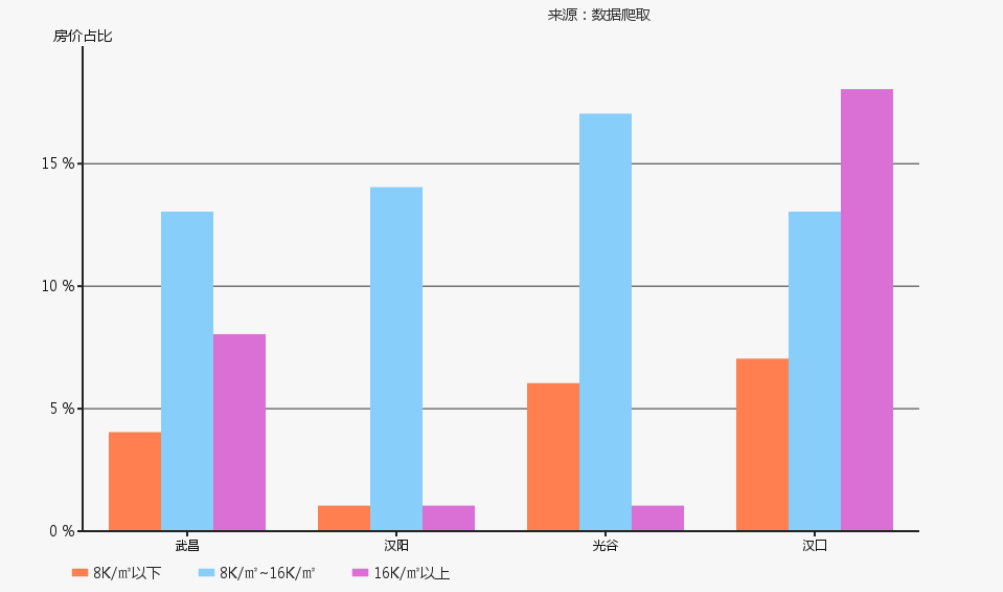

图 3-6　武汉房价分布图

表格 3-1 武汉各小区房产信息表

序号	字段	名称
1	Housename	房屋名称
2	totalprice	总价
3	housetype	户型
4	buildarea	建筑面积
5	unitprice	单价
6	faceat	朝向
7	floors	楼层
8	community_name	小区名称
9	detailaddress	详细地址
10	buildepoch	建筑年代
11	living_type	住宅类型
12	cmm_part_bool	人车分流
13	cmm_building_num	总楼栋数
14	cmm_totalpeople	总户数

根据现有的信息，我们爬取了武汉市各个小区的主要信息，其中包括房屋名称，房屋总价，房屋户型，房屋建筑面积，房屋单价，房屋朝向，房屋楼层，小区名称，小区详细地址，小区建筑年代，小区住宅类型，小区内是否人车分流，小区内总楼栋数，小区内居住的总户数等基本信息。通过小区的分布情况和房价可以建立房价和区域之间的映射关系，进而通过房价可以刻画区域的广告价值。

同时还对爬取的数据进行了分类，主要对地理要素 POI 分了 12 个大类。他们分别是交通设施（23131）、休闲娱乐（4352）、医疗（123）、房地产（435622）、政府机构（324）、教育培训（7534）、文化传媒（2435）、旅游景点（2453）、生

活服务（235464）、美食（39845）、购物（32325）、酒店（2324）。

表格 3-2 爬取到的 POI 网络数据分类

序号	分类	数据条数
1	交通设施	23131
2	休闲娱乐	4352
3	医疗	123
4	房产	435622
5	政府机构	324
6	教育培训	7534
7	文化传媒	2435
8	旅游景点	2453
9	生活服务	235464
10	美食	39845
11	购物	32325
12	酒店	2324

从数据中可以看到，房产数据、生活服务、交通设施和购物所占比重较大，这些地方正是广告投放地比较集中的地方，同时也是受众人群聚集之地。

3.5 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。

用户画像是建立在大量的用户数据基础之上的建模过程，整个用户模型是一个整体，而个体则是用户的标签（Tag），每一个标签都代表了观察、认识和描述用户的一个角度。本文就是通过标签-模型这种方法将用户的特征描述出来，其结构示意图如图 3-7 所示。

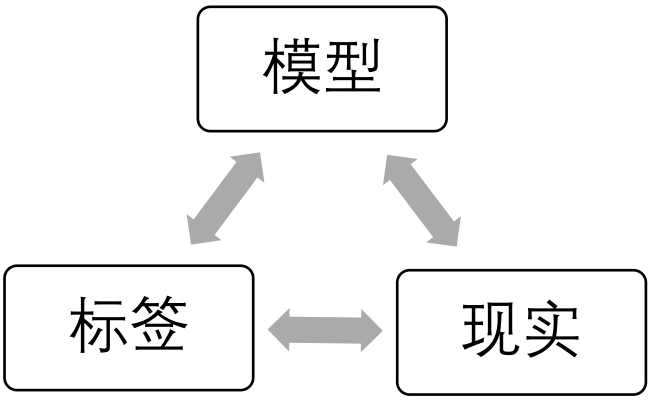

图 3-7 标签模型示意图

首先，根据业务需求定制类目标签体系，其次就是准备训练的数据。这些数据既包含用户的静态数据（用户注册信息等）也包含用户的行为数据（点击、广告屏蔽反馈等），都通过 UI 系统获取并通过日志系统存储在数据库中。收据收集之后，就需要进行数据清洗，因为数据中可能包含了一些空缺的或者暂时使用不到的部分。为了保证后期用户建模的准确性，避免结果的不准确，在数据挖掘之前应当对数据进行清洗，保留有效字段。

然后是数据标准化，这里通常指代的是夸终端数据整合，建立统一的标准，数据标准化之后就应该开始用户建模。常见的用户模型表示方法有：关键词集合表示法，使用用户的兴趣特征词表示用户模型。举个例如，用户 A 的爱好是健身，A 的模型就可以用{耐克、健身房、蛋白粉，瑜伽}表示。基于神经网络的表示法，使用特征化的网络状态模拟用户模型；最为流行的是基于向量空间模型的表示法。

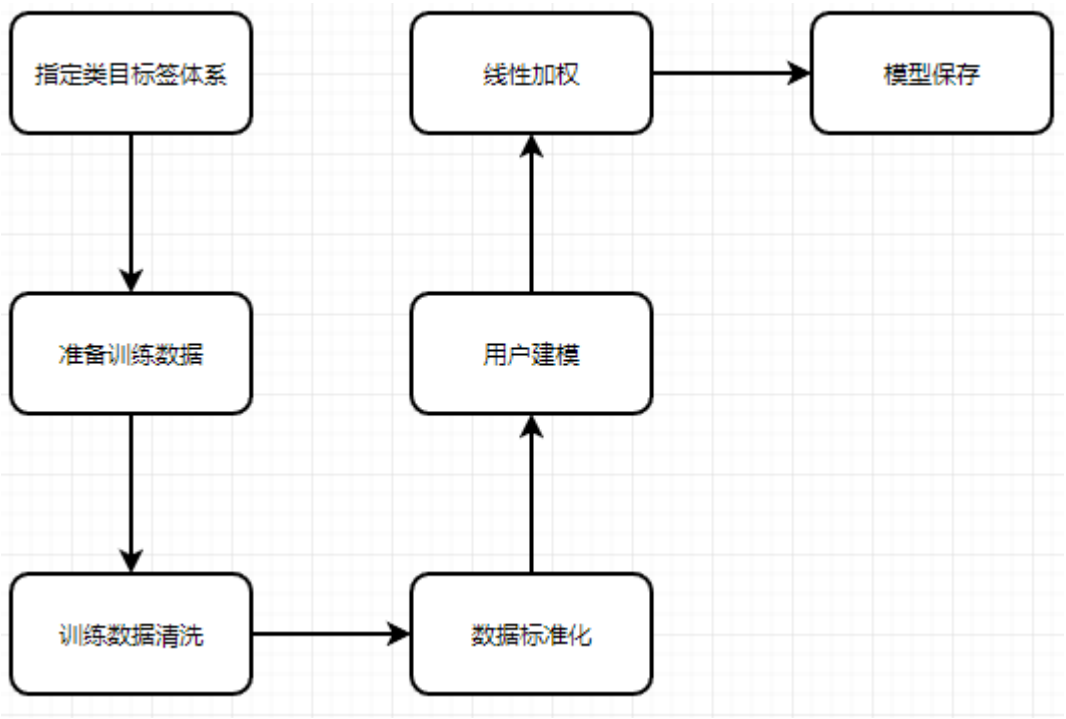

图 3-8 用户画像过程示意图

建立用户模型的方法有很多，在应用领域使用最多的主要有三种，手工定制建模、示例建模和自动建模。手工定制模型需要用户在兴趣选择界面勾选或输入文字内容来确定用户模型。例如手机新闻 App 客户端在用户首次使用时会要求用户选择感兴趣的新闻主题，而系统会依据用户选择的主题对用户进行新闻推送。但这种方法过多的依赖用户，会造成用户积极性下降，而且用户对自己不知道或者没有接触过的领域很难确定是否喜欢，进而导致用户模型不准确。示例用户建模则准确的来讲是一种根据枚举和用户反馈的建模方法，首先随机推动给用户信息，用户满意则点击，不满意就会否定。排除了用户不喜欢的，剩下的自然就是喜欢的类型。例如新浪微博 App 中插入的广告，通常在右上角会有是否满意的选项，如果不满意可以选择关闭。这种方法同样由于频繁的交互，降低用户的积极性，而且获得的示例包含较多的噪声样本和无效样本。而本文选择使用的是自动用户建模技术。

自动建模是指根据用户的行为信息之在后台进行数据挖掘然后建立模型，用户的参与度为 0。用户画像的需要的数据主要分为三类，第一类是自然数据，主要是用户的年龄、性别、住址等一些可以通过注册收集到的信息。这些信息文本自成标签，在实际提取中主要进行清洗工作。第二类是行为数据，包括用户的访问次数、访问停留时间、访问频度、转发与点赞等一些交互行为数据。这些行为

是离散的，需要通过数据分析，来建立用户的行为定向模型。第三类数据是用户访问的内容数据，主要是用户浏览的 web 内容等。同样需要进行文本分析来建立用户的模型。隐式反馈就是将用户的第二类和第三类数据记录到数据库中，通过将这些数据文本化，进行文本分析，进而得到用户模型。

3.6 本章小结

本章基于爬虫等数据挖掘的方法，爬取了信息房产和用户的相关信息，并对数据和信息进行本章首先介绍了系统的体系结构，给出了系统的总体框架。然后从各个模块出发介绍了各个模块的功能，模块算法的大致的步骤。在爬虫的组成中，网页解析器是基础模块，它处理网页的内容，提取新的 URL，处理标签等，所以在模块详解中，着重介绍了网页分析模块的组成及工作流程等。由于聚焦爬虫是面向主题定制的，本文研究的聚焦爬虫又侧重于主题数据的抽取，所以它相较于通用爬虫多出的主题定制和匹配模块是重点模块，因此，本章对主题定制以及主题匹配的过程也作了详细的介绍。

第四章 广告精确投放算法与冷启动

4.1 引言

通过第三章挖掘的地缘特征数据和受众群体数据特征，来分别计算算法和数据处理，基本步骤由数据采集，数据清洗，数据量化，然后通过算法计算出地域的广告投放价值和人群购买力价值。构建广告投放精准推荐系统相关算法，算法的核心是协同过滤算法，同时还考虑到推荐系统中会遇到冷启动问题，本文采取利用历史数据、利用用户偏好和设置问答的方式来解决广告推荐中的冷启动问题。

4.2 数据标引流程

利用城市大数据多元融合来对地段人群进行区分在当代极具价值，它可以有效衡量一个地段的整体特征。在纽约时代广场的女性必然大概率会比来自偏远山区的女性更能消费得起高档化妆品，这种地缘特征也为广告投放商提供了参考，这也是在不同地段见到不同广告的根本原因。

然而传统的广告投放方式具有太多不确定性，如地段内的年龄分布、阶层分布均是无法直接获取的数据，而本算法目标便是通过当前大数据处理方案，通过机器学习的方式挖掘出导致不确定性的本质特征，提供稳定的地段特征信息，对地段内人群的基本购买力进行划分，从而实现广告的精准投放。

4.2.1 算法概述

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。

4.2.2 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。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。

表格 4-1 所需要数据列表

序号	数据类型	性质
1	连续时间短的的房价信息（用于数据标引）	必需数据
2	路网数据（道路，车站，地铁）	必需数据
3	河流，山脉等地理数据	必需数据
4	POI 数据（商场，学校等）	必需数据
5	人流数据（滴滴，公交）	非必需数据
6	PM2.5 数据（间接衡量交通、工业发展程度）	非必需数据

4.2.3 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） 为数据项标引

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

图 4-1 量化河流示意图

根据图示，可以看出，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。

图 4-2 量化公交线路示意图

根据图示，可以看出，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

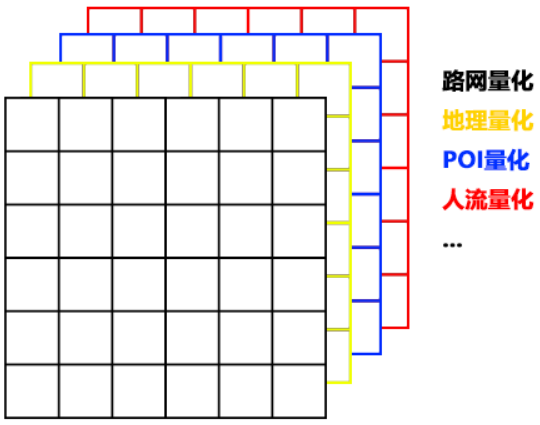

图 4-3 标记样本数据集组合

（5） 得到拥有一组样本的训练集

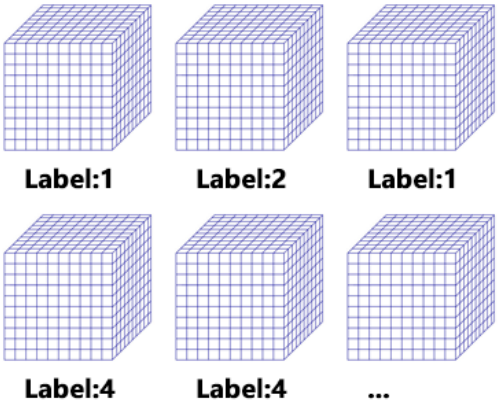

图 4-4 得到样本的训练集和测试集

4.2.4 总体结构

总体结构为：

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。

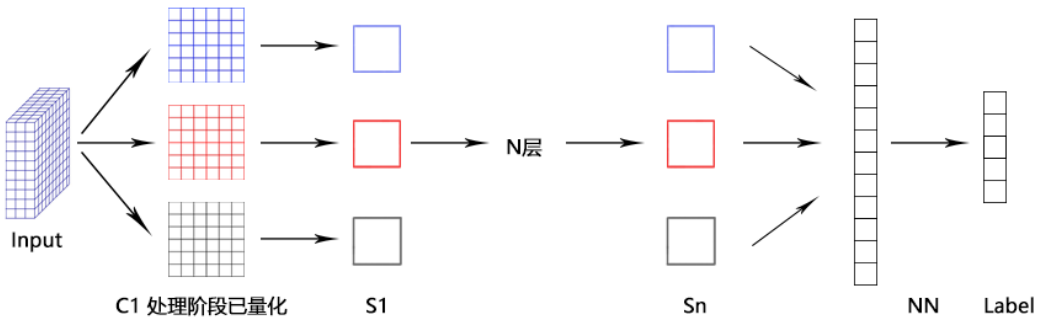

图 4-5 算法整体结构

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

4.2.5 计算结果

计算得到全城的地段价值分类，可以显著的发现最具购买力的地段。

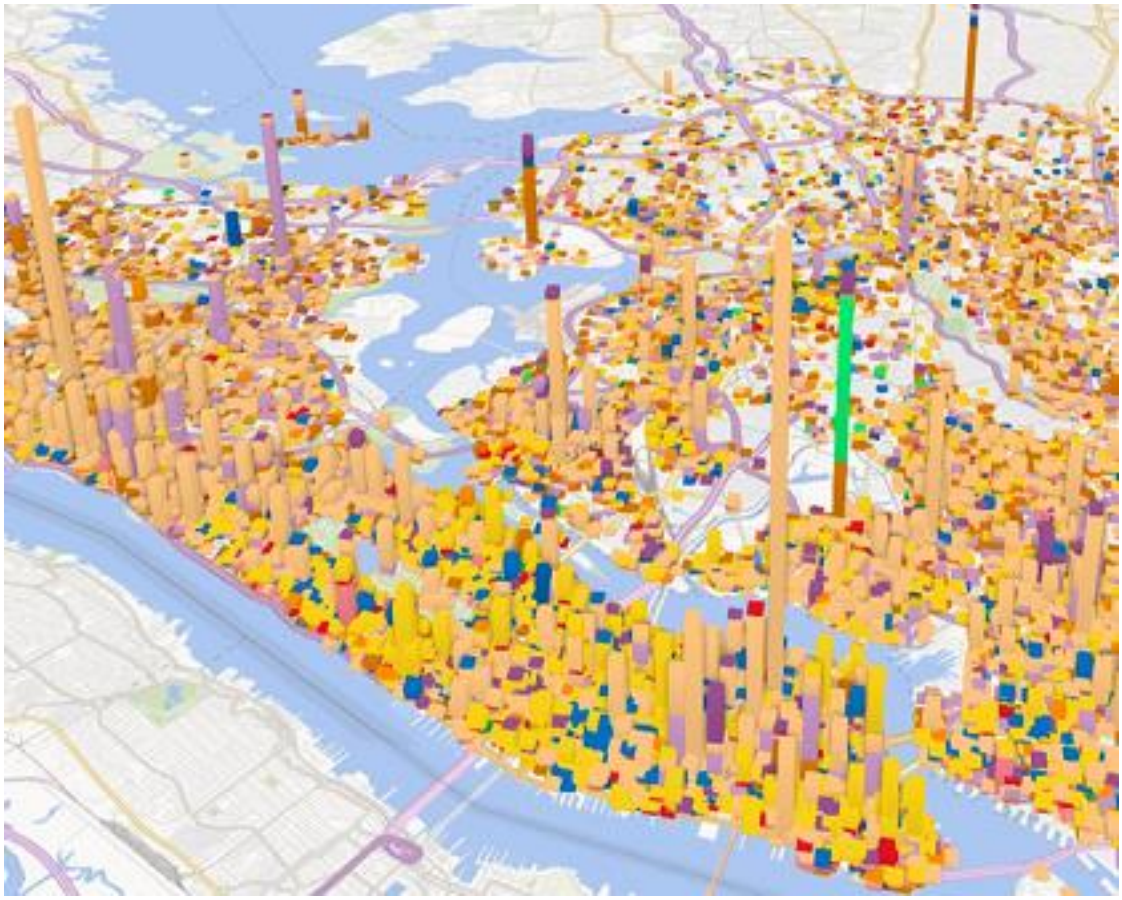

图 4-4 全城购买力和地段价值分布图

从图 4-6 中可以看到，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。

4.3 广告投放推荐模型

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。

4.3.1 核心问题解决方案与算法

（1） 算法目标：

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。

（2） 已达成的前置条件：

与武汉市广告生产投放中介的合作

（3） 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

（4） 可直接提取的特征

a. 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。

b. 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。

（5） 可根据服务计算理论建立模型提取的特征

a. 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。

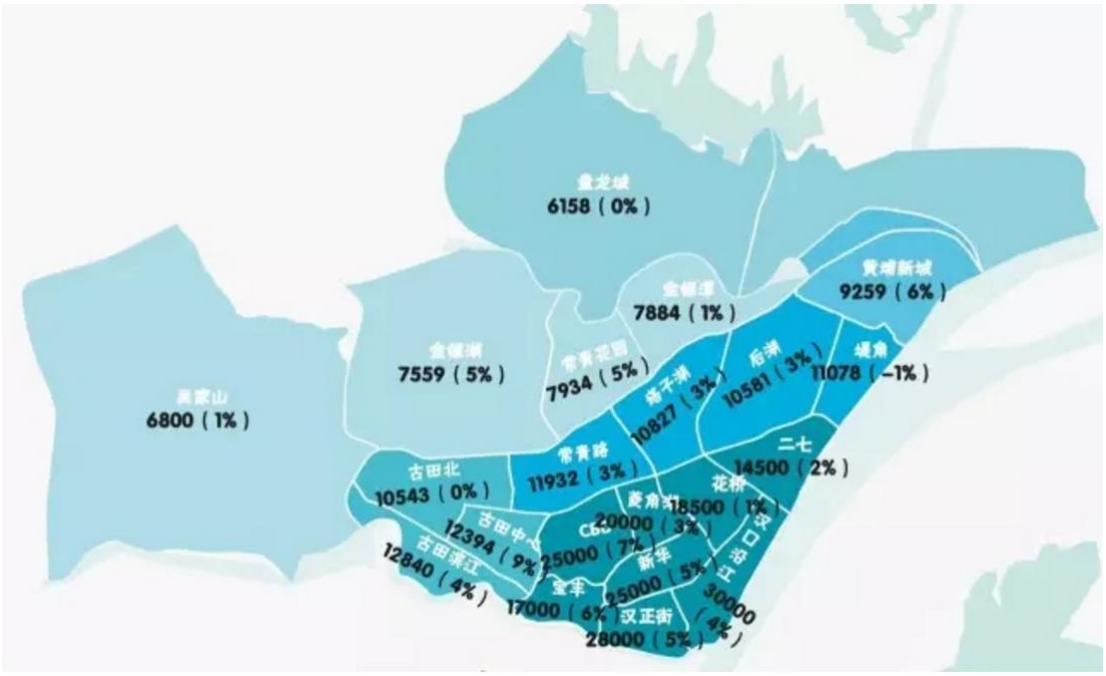

图 4-5 武汉汉口房价分布情况图

图 4-5 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。

b. 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：

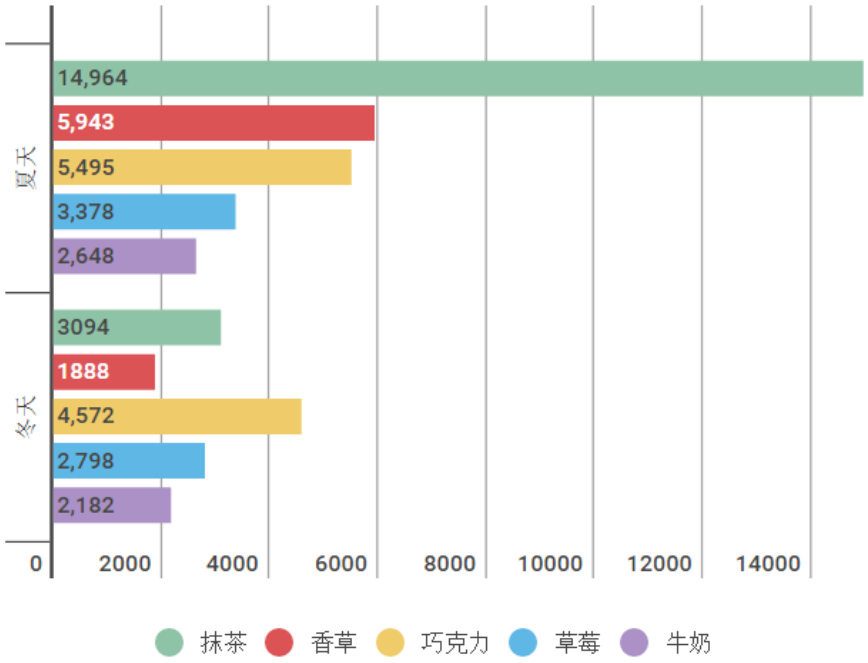

图 4-6 冰淇淋广告投放商的时间-销量图

从图 4-6 可以看出，冰淇淋广告投放商在冬季的投放需求明显减少。

（6） 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：

a) 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。

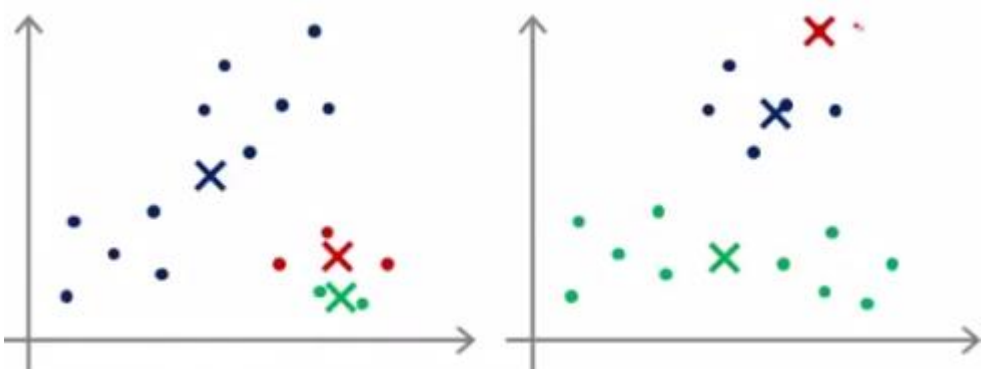

图 4-7 算法优化回归图

b）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。

c）建立简要 BP 神经网络（图 4-8），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。

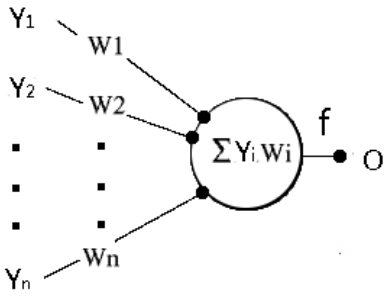

图 4-8 BP 神经网络节点图

d）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

（7） 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 4-8 所示的广告关系模型，进一步挖掘影响广告投放的因素。

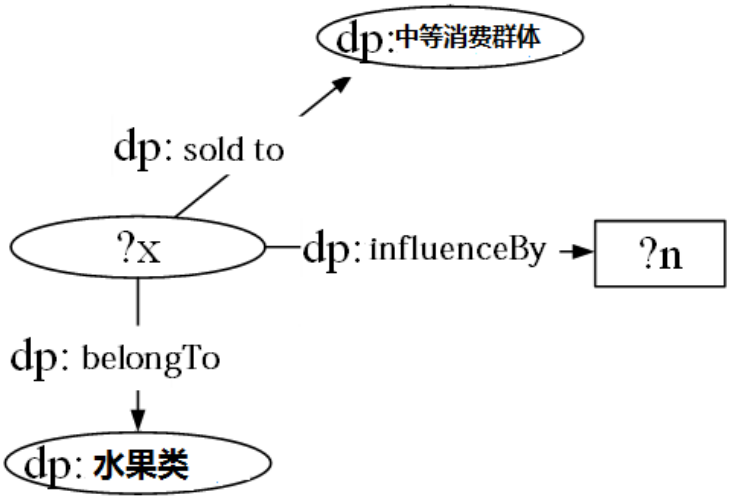

图 4-8 广告关系模型图谱

4.3.2 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：能够过滤难以进行机器自动基于内容分析的信息。能够基于一些复杂的难以表达的概念（用户特征）进行过滤。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。

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。

4.4 冷启动系统

推荐系统冷启动在不同的产品，不同的应用场景会有不同的做法，通用的做法其他的知友都已经有说明，即通过不同的维度获取用户的基本特征，操作习惯，从而进行粗粒度的推荐，但确实说起来容易做起来难。因为在中国，只有几家比较大的如腾讯，阿里，百度掌握着用户的行为数据，小公司做起来有点知易行难的感觉，在这里分别说说自己的一些思路和方法。

4.4.1 利用用户历史数据

利用用户在其他地方已经沉淀的数据进行冷启动。在腾讯等大公司的产品，确实是可以通过各大产品打通的日志系统，提取用户的行为特征去确定用户是个什么样的人。比如现在 QQ 音乐的猜你喜欢电台想要去猜测还没有用过 QQ 音乐用户的口味偏好，一大优势在于可以利用其他腾讯平台的数据，比如在 QQ 空间关注了谁，在腾讯微博关注了谁，这些都可以作为推荐系统的冷启动数据，甚至进一步，比如在腾讯视频刚看了一部很火的动漫，如果在 QQ 音乐推荐了一首这个动漫的主题曲，你是否会觉得很惊喜呢？=P 所以，在这方面可以做的尝试，就是获取用户在其他平台已有的数据。题主担心的是一个初创网站或 app 用户注册前还没有他的数据表现，不妨尝试将注册路径改为用新浪/QQ/微信等社交平台登录，一方面可以降低用户注册成本提高转化率，一方面可以同时获得用户的社交信息，从而获得推荐系统的冷启动数据。举个大家都应该知道的产品——“今日头条”，号称 5 秒钟知道你的兴趣偏好，其实也是在用户登录新浪等社交平台后，获取用户的关注列表，以及爬取用户最近参与互动的 feed（转发/评论/赞）进行语义分析，从而获取用户的偏好。这种方法无论公司或平台大小，其实都可以尝试，会比盲目的热门推荐效果会好。

4.4.2 利用用户兴趣偏好

利用用户的手机等兴趣偏好进行冷启动。Android 手机开放度较高，因此对于各大厂商来说多了很多了解用户的机会，就是——用户除了安装的应用之外，还安装了其他什么应用。举个例子，当一个用户安装了美丽说，蘑菇街，辣妈帮，大姨妈等应用，是否就是基本判定该手机用户是个女性，且更加可以细分的知道是在备孕还是少女，而安装了 rosi 写真，1024 客户端带有屌丝气质的应用则可以锁定用户是个屌丝，此时对于应用方来说，是一个非常珍贵的资源。比如一个新闻应用如今日头条，拿到了这些用户安装应用的数据，用户首次安装就可以获得相对精准的推荐，不明真相的用户还会暗赞我靠这应用这么符合我口味！目前读取用户安装的应用不仅是 APP 应用商店的标配，新闻类，视频类做数据推荐的应用也有一些开始读取这块的数据，这个对于冷启动是相当有帮助的。当然，这种数据也要为用户做好保密和数据加密。另外如豌豆荚锁屏，360 卫士 app 更是做了检测用户每天开启应用的频率等等，这种相比只了解用户安装什么应用，对用户的近期行为画像会更为精准。

4.4.3 通过选项采集初始信息

制造选项，让用户选择自己感兴趣的点后，即时生成粗粒度的推荐。相对前面两个来说，路径不够自然，用户体验相对较差，但是给予足够好的设计，还是能吸引用户去选择自己感兴趣的点，提升转化率。比如网易云音乐的私人 FM，由于没有其他用户行为数据，做口味测试则变得很重要了。而简单幽默的文案引导加上简单的几个选择，也不失为一个好的冷启动方法。

4.4.4广告系统冷启动

相似人群拓展的目的是基于广告主提供的目标人群，从海量的人群中找出和目标人群相似的其他人群。在实际广告业务应用场景中，相似人群拓展能基于广告主已有的消费者，找出和已有消费者相似的潜在消费者，以此有效帮助广告主

挖掘新客、拓展业务。相似人群拓展基于广告主提供的一个基础特征人群，自动计算出与之相似的人群（称为扩展人群）。

冷启动特征主要包括广告特征和群体特征，其中广告特征有：广告的基本信息（广告名称、投放时间等）、广告的推广目标、广告的标签、投放平台、投放的广告规格、所投放的广告创意、广告的受众、广告出价信息、目标购买力、目标年龄层、目标 POI；群体特征：地段购买力、段模糊年龄层、地段周边 POI。利用广告特征和群体特征进行相似性计算。推荐系统需要数据作为支撑。但在刚刚开始做推荐的时候，是没有大量且有效的用户行为数据。这时候就会面临着“冷启动”的问题。没有用户行为数据，就利用商品本身的内容数据。这将是算法前期的做法。

根据广告与群体之间的共同特征为维度，建立一个多维的空间，对单一维度上的评价组成的坐标系即可定位群体在这个多维度空间中的位置，那么任意两个位置之间的欧式距离在一定程度上反应了广告与群体之间的相似程度。

就其意义而言，欧氏距离越小，广告与投放群体相似度就越大，欧氏距离越大，广告与投放群体相似度就越小。

4.5 本章小结

本章主要讨论了数据采集，数据清洗，数据量化，然后通过算法计算出地域的广告投放价值和人群购买力价值。构建广告投放精准推荐系统相关算法，算法的核心是协同过滤算法，同时还考虑到推荐系统中会遇到冷启动问题，本文采取利用历史数据、利用用户偏好和设置问答的方式来解决广告推荐中的冷启动问题。

第五章 系统的实现与实验评估

5.1 引言

基于前面几个章节的数据采集、数据分析、算法设计等流程，构建起了一套基于机器学习和大数据分析技术的广告投放精准推荐体系。本章将从需求分析，分别分析市场需求和用户需求，然后进行系统设计，逐步完成系统的开发，最终对比系统使用后结果的提升和评估结果。

5.2 需求分析和系统设计

5.2.1 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。

5.2.2 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。

5.3 系统架构

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。

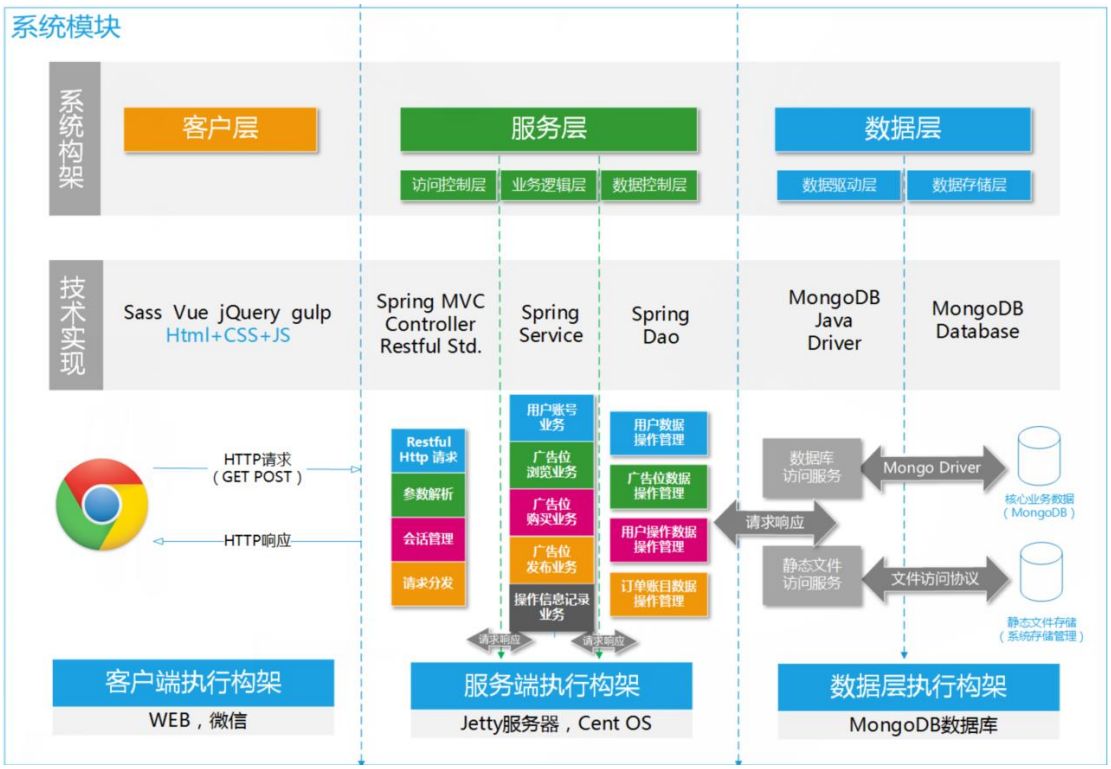

图 5-1 系统模块图

系统模块如图 5-1 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。

5.3.1 系统架构

基于大数据分析和用户行为的广告投放精准推荐系统原型系统，在进行系统总体架构设计时主要从以下几个方面来进行，网站架构、软件架构、系统流程。

（1）网站架构

网站的架构设计是为了更好的构建和管理一个通信网络，提供架构和技术基础的规划。包括用户使用的网络协议、接口类型和可食用的网络布线等。

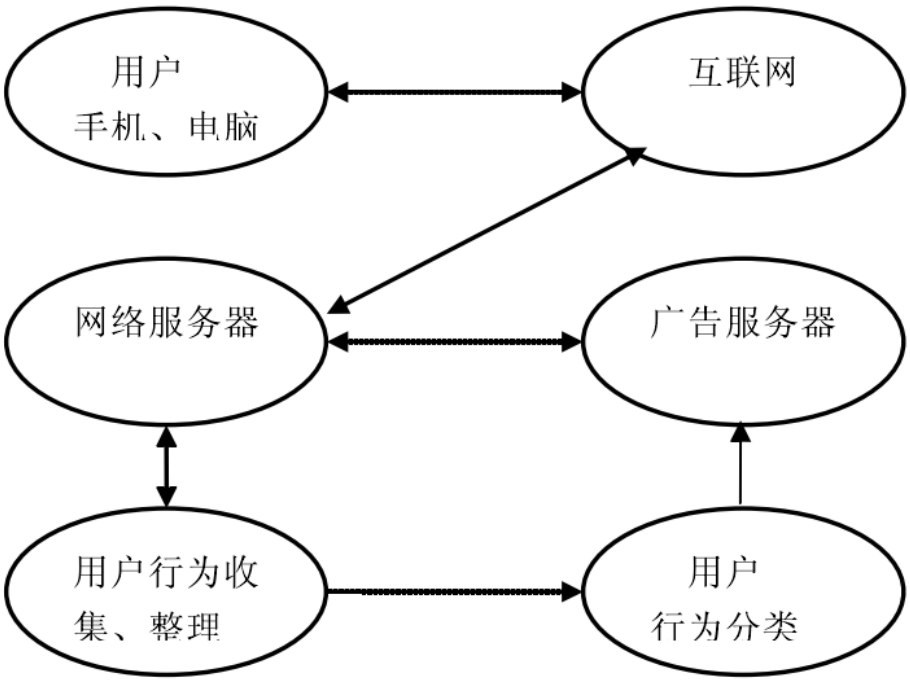

图 5-2 系统架构图

从图 5-2 我们可以看出，用户通过使用电脑，智能移动终端（手机、平板电脑）等登录网站，通过互联网与网络服务器建立连接，就可以浏览和获取自己的所需要的网页资源。网络服务器在中间提供桥梁和媒介作用，一方面可以响应用户的请求，让用户得到所需要的内容和咨询，另一方面，反过来可以收集到用户的信息和行为，然后反馈到服务器。后台根据采集到的用户的行为信息和数据对用户数据进行整理和分类，通过数据挖掘的方法分析得到用户的画像，同时还会从广告服务器上获取到与之相对应的广告，并且展示在网页端展示给用户。手机用户行为和整理模块对用户信息的静态行为和聚类分析的动态信息，每个用户每次上网浏览多个网页内容，也会在网上产生大量的操作，比如在淘宝上在线购物行为，点击立即购买，点击支付宝水费基金电费，在平安保险公司网站购买保险。在这个过程中，用户浏览或者点击网页的主题和关键词信息，细分用户浏览网页

的行为，建立用户行为收集和排序模块。使用相应算法对所有用户行为主题进行评分，我们得到用户行为关键字或主题评分列表。根据用户行为中每个关键词的得分，广告服务器对所有关键词或主题进行排序，得到用户行为最匹配的关键词信息。最后，将关键字信息的广告提供给用户以完成准确的广告。

（2）软件架构

基于数据挖掘的精准广告原型系统是由 JAVAweb 开发的一个系统。三层架构用于开发过程。三层体系结构通常包括逻辑层，接口层和数据层。三层架构可以实现高内聚和低耦合效果。1，逻辑层用于控制数据层，实现数据逻辑处理。2，界面层是用户输入数据和数据采集的接口，界面层也可以保证信息的机密性。3，数据层，由该层完成的事务直接操纵数据库，针对数据的添加，删除，更新，搜索等。三层架构的特点是：（1）具有很强的可扩展性;一个层次的人只能在这个层次上工作。（2）开发人员只需关注系统结构中的某一层;（3）新实施方便与原有水平交换;（4）可以标准化;（5）每层重用都很容易;（6）减少层与层之间的耦合;（7）项目清晰，未来的维护和升级方便;（8）安全性高。用户只能通过逻辑层访问数据层，封装了大量危险功能。

控制层充当桥接器，位于接口层和数据层之间。 用户界面层，请求控制层从数据层中提取数据，并将相应的数据提供给接口层。 界面层是用户可以看到的界面。 用户输入数据并获取界面层中的数据。 界面层也具有安全性，以防止不应该被看到的人看到机密信息。 数据层的目的是确保数据的完整性和安全性。用户的请求通过逻辑层来访问数据。 它通常依赖于数据库服务器，如 Mysql 和 Oracle 等。

5.3.1 开发环境介绍

系统开发环境配置介绍：随着云计算的发展，阿里巴巴、亚马逊等提供可以供用户使用的云主机，根据项目的需要，我们这里选取阿里云和百度云的作为主机服务器，服务器的作用是部署网站系统，搭建数据库环境，统计分析用户行为，如表 5-1 所示。

表 5-1 系统云服务器配置

服务器	功能描述
阿里云服务器	系统运行平台，用于为广告商和用户提供服务，同时监测用户行为
百度云服务器	构建数据库存储环境，用于存储、整理分析需求和用户行为

软件环境如表 5-2 所示，描述了系统构建过程中所用到的运行环境和软件配置情况。运行的系统是 Linux CentOS 7，利用 JDK 的版本是 JDK 1.8 来配置 Java 开发环境，运用 Mysql5.6 和 MongoDB 来进行数据的保存，开发工具选择 JetBrains IDEA 来搭建开发环境和开发编译器,Java Web 的框架利用的 Spring + Spring MVC + Mybatis。

表 5-2 环境和软件配置

名称	版本	功能
系统平台	CentOS 7	系统环境
JDK	JDK 1.8	Java 开发环境
Mysql	Mysql5.6	保存关系型数据
MongoDB	MongDB	分布式数据库
JetBrains IDEA	2017.05	Java 开发编辑器
SSM	Spring 4 Spring MVC Mybatis 5	开发框架

5.3.2 数据库设计

基于用户行为的精准广告投放系统主要包括广告投放商、广告、广告类别、爱好等类型,数据库模型如下图所示，这三张表表间关系如下所示。

表 5-3 广告表（ad_table）

备注	类型	长度	约束	其他
Ad_Id（广告 ID）	Int	100	主键	非空
Ad_title（广告标题）	Varchar	100	否	可以为空
Ad_content（广告内容）	Varchar	255	否	可以为空
Ad_agentId(广告投放商)	Int	100	主键	非空
Ad_typeId(广告类别)	Int	100	主键	非空

表 5-4 用户表（User_table）

备注	类型	长度	约束	其他
User_Id(用户 ID)	Int	100	主键	非空
Ad_Id（广告 ID）	Int	100	主键	非空
Joy_Id（喜好 ID）	Int	100	主键	非空
User_name（用户名）	Varchar	100	否	可以为空
User_address（用户地址）	Varchar	255	否	可以为空
User_gender（用户性别）	Varchar	20	否	可以为空
User_phone（用户电话）	Varchar	50	否	可以为空
User_createTime（创建时间）	Varchar	50	否	可以为空

表 5-5 广告类别表（ad_type_table）

备注	类型	长度	约束	其他
Ad_typeID(广告类别 ID)	Int	100	主键	非空
Ad_type(广告类别)	Varchar	100	否	可以为空

5.3.3 系统原型界面展示

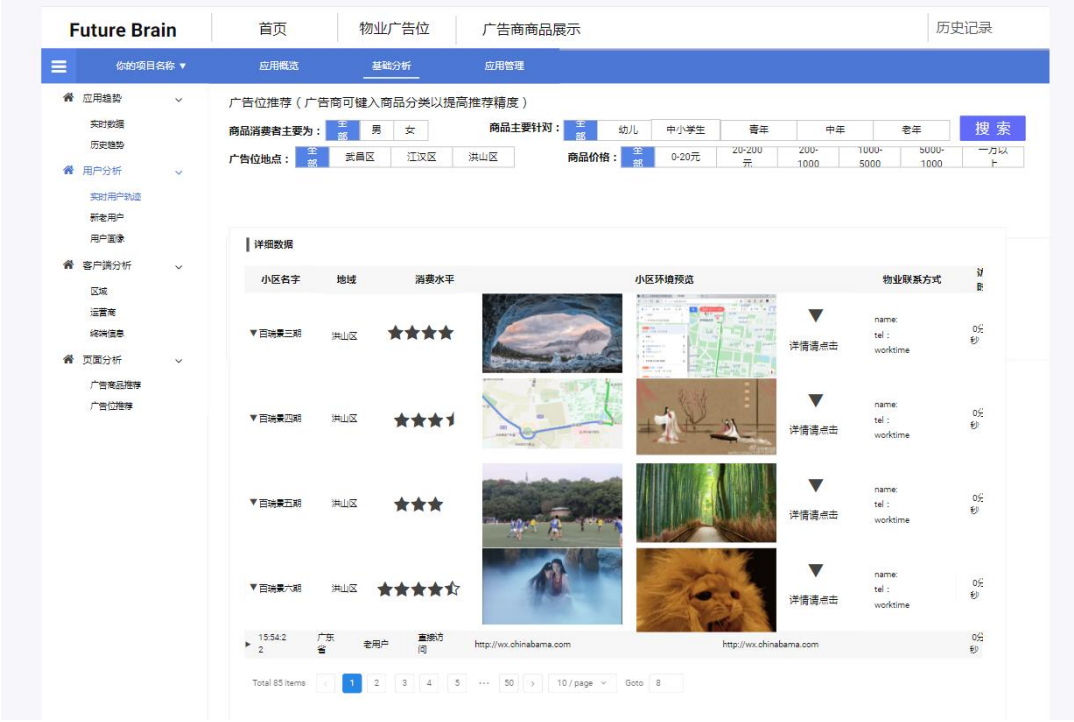

图 5-3 广告推荐界面展示图

图 5-4 用户画像分析界面图

5.4 实验结果评估

准确的广告宣传平台的数据主要是离散数据，因此我们在数据挖掘方法中使用偏差分类算法对用户的行为进行分类。用户行为的偏差分类过程是为每个用户计算每个行为数据分为每个类别的概率，行为数据的概率最大的类别是用户的最终类别。

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。

表 5-6 精准广告营销与传统广告营销点击率比较

广告名称	投放次数	传统营销	传统营销	传统营销	精准营销	精准营销	精准营销
		显示次数	点击次数	点击率	显示次数	点击次数	点击率
车险	4000000	4000000	10050	0.25%	3405212	17101	0.52%
银行业务	4000000	4000000	13501	0.33%	3506002	19012	0.56%
寿险	4000000	4000000	11012	0.27%	2890062	13811	0.51%
证券	4000000	4000000	9000	0.24%	2780753	13221	0.49%
信托	4000000	4000000	8903	0.21%	2478617	11031	0.45%
期货	4000000	4000000	11000	0.28%	3441089	19115	0.57%
基金	4000000	4000000	9500	0.24%	3000691	13117	0.46%
资产管理	4000000	4000000	12032	0.27%	3085135	15461	0.51%
租赁	4000000	4000000	13003	0.48%	3164089	15494	0.62%

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点击率大多了。

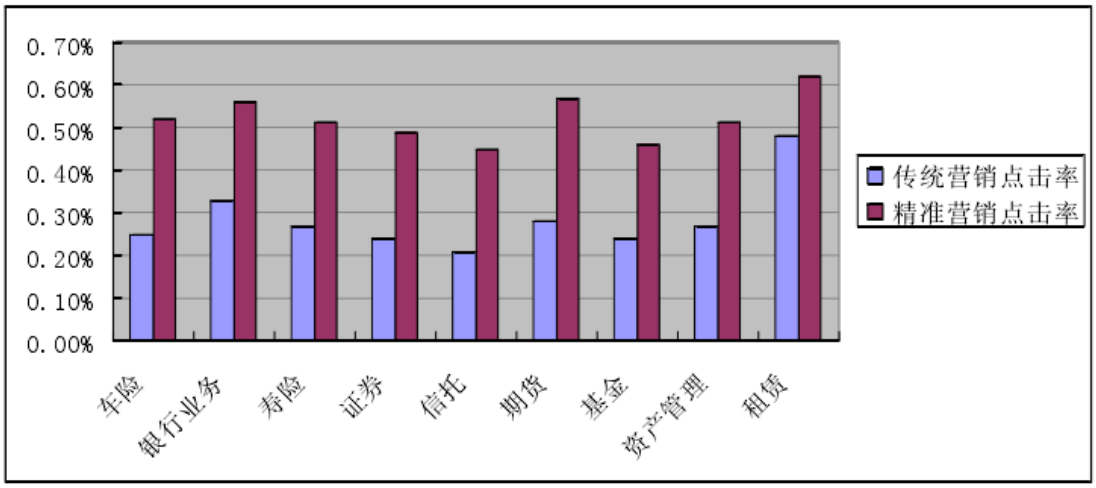

图 5-5 实验结果对比图

通过实验的结果的对比图可以发现，精准营销的点击率比传统营销的点击率有大幅度的提高，推荐的效率好。

第六章 总结与展望

6.1 总结

本文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。

本推荐系统共有三个模块构成，即冷启动数据系统模块、广告位推荐系统模块、人物画像系统模块。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。

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。

6.2 展望

推荐算法的准确性与数据的质量密切相关，所以要不断提高数据的质量，使得算法的精度得以提高。同时在广告商的投放策略中，不仅仅受到本文提出的因素的制约，还有更多的要素未能体现在系统中。系统所采取的协同过滤算法是推荐系统中较为基本的算法，虽然一定程度上提高了推荐的精度，但是仍然要挖掘更有效的算法，使得推荐的结果更加准群。

对于用户行为的数据采集工作仅仅局限于线上，不够全面，有些数据的采集也可能不够客观，要不断的优化系统，使得系统对用户的数据采集效率更好，内容更多，更全面。系统的构建虽然可以一定程度上实现“去库存化”的站略目标，但是建立的广告投放模型仍然具有一定的风险，还需要采用更加专业的风险评估模型来进一步的对系统进行评估，为社会做出更得贡献。

冷启动系统虽然可以给系统提供推荐的基础，但是数据的更新合同部同样非常重要，随着系统的使用，应该使用户的历史数据不断地补充进冷启动系统内，让系统不断地进化和发展，与时俱进，只有这样才能够促进系统的智能化。随着机器学习和人工智能的发展，深度学习的算法也被逐渐广泛应用于推荐领域，他在很多领域的成功，必然会促进广告精准投放行业的发展和进步。

参考文献

[1]. 周立柱，林玲.聚焦爬虫技术研究综述[J].计算机应用，2005,25(9):1965-1969.

[2]. 齐保元.网络爬虫[DB/OL]. http://baike.baidu.com/view/ 284853.htm, 2006-1 -10/2011-4-20.

[3]. Lefteris Kozanidis. An Ontology-Based Focused Crawler[J]. Computer Science, 2008, 5038: 376–379.

[4]. M.Yuvarani，N.Ch.S.N.Iyengar，A.Kannan. LSCrawler: A Framework for an Enhanced Focused Web Crawler based on Link Semantics[A]. The 2006 IEEE/WIC/ACM International Conference[C]. Washington, CS Press,2006.794-800.

[5]. Wenxian Wang, Xingshu Chen, Yongbin Zou. A focused crawler based on naive Bayes classifier[A]. Third International Symposium on Intelligent Information Technology and Security Informatics[C].Washington, CS Press,2010.517-521.

[6]. Yulian Zhang，Chunxia Yin，Fuyong Yuan. An application of improved PageRank in focused crawler[A].Fourth International Conference on Fuzzy Systems and Knowledge Discovery (FSKD 2007)[C]. Washington, CS Press,2007.331-335.

[7]. Rob Miller. Websphinx[DB/OL]. http://www.cs.cmu.edu/~rcm/websphinx, 2002-7 -8/2011-4-20.

[8]. Li Peng，Teng Wenda. A focused web crawler face stock information of financial field[A]. Intelligent Computing and Intelligent Systems (ICIS)[C]. Newyork:IEEE,2010.512-516.

[9]. 周建梁.聚焦爬虫原理及关键技术研究[J].科技资讯，2008,(22):26.

[10]. 林海霞，司海峰，张微微. 基于 Java 技术的主题网络爬虫的研究与实现[J].微型电脑应用，2009,25(2): 56-58.

[11]. 语法分析.http://baike.baidu.com/view/487037.htm#sub487037, 2011-1-20/2011 -4-20.

[12]. 姜梦稚.基于 Java 的多线程网络爬虫设计与实现[J].微型电脑应用，2010,26(7): 56-58.

[13]. 蔡欣宝，郭若飞，赵朋朋，崔志明. Web 论坛数据源增量爬虫的[J].计算机工程. 2010.5,36(9)：285-287.

[14]. 张丽敏.垂直搜索引擎的主题爬虫策略[J].电脑知识与技术,2010.5,6(15)：3962-3963.

[15]. 杨靖韬，陈会果.对网络爬虫技术的研究[J].科技创业月刊，2010,(10):170-171.

[16]. 刘磊安，符志强.基于 Lucene.net 网络爬虫的设计与实现[J].电脑知识与技术，2010.3,6(8)：1870-1871，1878.

[17]. 叶勤勇.基于 URL 规则的聚焦爬虫及其应用[D].浙江：浙江大学硕士学位论文，2007.5.

[18]. 夏道勋，谢晓尧.基于 Web 的专用爬虫的研究[J].贵州师范大学学报（自然科学版），2009.8，27（3）：92-95.

[19]. 杨志伟，王鑫.基于本体的气象领域聚焦爬虫[J].中国管理信息化，2011.2,14(4)：60-62.

[20]. Pedro Huitema, Perry Fizzano. A Crawler for Local Search[A]. 2010 Fourth International Conference on Digital Society[C]. New York:IEEE, 2010.86-91.

[21]. Qing Gao, Bo Xiao, Zhiqing Lin, Xiyao Chen, Bing Zhou. A High-Precision Forum Crawler Based on Vertical Crawler[A]. IEEE International Conference on Network Infrastructure and Digital Content[C]. New York:IEEE, 2009.362-367.

[22]. 孙立伟，何国辉，吴礼发.网络爬虫技术的研究[J].电脑知识与技术, 2010.3,6(15)：4112-4115.

[23]. 张成奇.支持 Ajax 的 Deep Web 爬虫设计与实现[D].上海：上海交通大学工程硕士论文，2009.

[24]. 罗兵.支持 AJAX 的互联网搜索引擎爬虫设计与实现[D].杭州：浙江大学硕士学位论文，2007.

[25]. 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.

[26]. 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.

[27]. 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.

[28]. 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.

[29]. 邹海亮.可定制的聚焦网络爬虫[D].上海：东华大学硕士学位论文，2008.12.

[30]. 倪贤贵.聚焦爬虫技术研究[D].无锡：江南大学硕士学位论文，2008.04.

[31]. Li Wei-jiang,Ru Hua-suo,A New Algorithm of Topical Crawler[A]. 2009 Second International Workshop on Computer Science and Engineering[C]. New York:IEEE,2009.443-446.

[32]. Shan Lin,You-meng Li,Qing-cheng Li. Information Mining System Design and Implementation based on Web Crawler[A]. System of Systems Engineering[C]. New York:IEEE,2008.1-5.

[33]. Pooja gupta,Mrs.Kalpana Johari. Implementation of Web Crawler[A]. Second International Conference on Emerging Trends in Engineering and Technology [C]. New York :IEEE ,2009.838-843.

[34]. 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，2015，9：81-86.

[35]. 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.

[36]. 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.
[37]. 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.
[38]. 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.
[39]. 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.
[40]. 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.
[41]. 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.
[42]. 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.
[43]. 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.
[44]. 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.
[45]. 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.
[46]. 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.
[47]. 张敏，茹立云，马少平.网络检索用户行为可靠性分析[J].软件学报.2010，5：1055-1066.
[48]. 徐川，王娟，赵国锋.基于网络用户行为的网站发展研究[J].计算机应用研究.2014，4：1154-1165.
[49]. 郭岩.基于网络用户行为的搜索引擎系统 SISI[J].计算机工程.2014，8：9-12
[50]. 张彤，孙全忠，闫东升.基于日志分析的网络用户行为分类研究[J].广东公安科技.2015，7：28-31.
[51]. 张庆娟.新媒体背景下网络精准广告的投放研究[J].新闻研究.2015,11:188-199.
[52]. http://site.douban.com/widget/votes/2325584/8848/
[53]. Hagen P.R,Manning H,and Souza R.Smartpersonalization.2007,Forrester Research:Cambridge, MA.8-21.
[54]. Mobasher B., et al. Combing web usage and content miningfor more effective personalization. In the International Conference on E-commerce and Web Technologies (ECWEB2000).2000.Greenwich,UK.
[55]. Hyoseop Shin,Minsoo Lee, Eun Kim. Personalized digital TV content recommendation with integration of user behaviorprofiling and multimodal content rating[J]. IEEE Transactions on Consumer Electronics,2009:1417-1423.
[56]. Lee, Seunggwan, Lee Daeho, Lee Sungwon. Personalized DTVprogram recommendation system under a cloud computing environment[J].IEEE Transactions on Consumer Electronics,2010:1034-1042.
[57]. 张春阳，周继恩，蔡庆生.基于数据仓库的决策支持系统的建构计算机工程,2002,26(4):23-43.
[58]. 陈德军，陈绵云.基于数据仓库的 OLAP 在 DSS 中的应用研究[J].网络传播，

2003,23(1):30-31.
[59]. 王敏.顾客资产价值视角下的企业精准广告营销[J].企业经济，2010，173(5):27-31.
[60]. 陆南.基于互联网新媒体时代的精准广告营销研究[J].新闻知识,2014,15(9):127-136.

致　谢

漫漫人生路，求学十余载。转眼间，我在武汉大学已经度过了三个精彩而充实的春华秋实，回顾以往的学习和生活，感慨万千。短短几年间，我已经从一名不谙世事的珞珈少年，成长为自强弘毅求是拓新的武大青年。我有幸认识了很多学生渊博的老师和优秀的同学，你们是我求学生涯中最大的收获与财富。正是你们的帮助与陪伴，我才能顺利完成学业，为硕士阶段的学习画上一个句号。

“饮其流者怀其源，学其成时念吾师”。在本文完成之际，首先我要中心的感谢我的导师郑宏教授。师从郑宏老师以来，但我深为他那高深的学术造诣、严谨的治学态度、平易近人的师长风范所折服，也为他敏锐的学术眼光、儒雅的个人情怀、博大的胸襟抱负所惊叹。在与郑老师的每次交流中我都能获益良多；感谢李老师在我开题、中期及答辩过程中为我指出问题，并严格要求，提出中肯的意见。在恩师的无私帮助和鼓励下，毕业论文才得以顺利完成。恩师言传身教，使我铭感五内。