



中山大學
SUN YAT-SEN UNIVERSITY

本科生毕业论文（设计）

Undergraduate Graduation Thesis (Design)

题目 Title: 基于用户画像的广告后台系统
的设计与实现

院系
School (Department): 数据科学与计算机学院

专业
Major: 计算机系网络工程

学生姓名
Student Name: 陆荣志

学号
Student No.: 14348090

指导教师(职称)
Supervisor (Title): 周凡（教授）

时间: 2018 年 4 月 19 日

Date: April 14, 2018

毕业论文（设计）成绩评定记录

Grading Sheet of the Graduation Thesis (Design)

指导教师评语

Comments of Supervisor:

能够按时完成广告系统的需求分析、设计与实现。并且根据项目的情况，结合用户画像，对广告精准投放算法进行改进。对于系统的不同功能，能够考虑到负载均衡。如果能够有结果展示和系统测试就更好了。

成绩评定

Grade: 良

指导教师签名

Supervisor Signature :

周凡

Date:2018.4

答辩小组意见

Comments of the Defense Committee:

成绩评定

Grade:

签名:

Signatures of Committee Members

Date:

院系负责人意见

Comments of the Academic Chief of School:

成绩评定

Grade:

签名

Signature:

院系盖章

Stamp:

Date:

表一：毕业论文（设计）开题报告

Form 1: Research Proposal of Graduation Thesis (Design)

论文（设计）题目

Thesis (Design) Title: 基于用户画像的广告系统的设计与实现

（简述选题的目的、思路、方法、相关支持条件及进度安排等）

（Please briefly state the research objective, research methodology, research procedure and research schedule in this part.）

通过分析广告后台系统的需求，设计出一个便于管理的广告系统。并且广告能够精准、定点去投放。

广告后台使用 WEB 服务器进行搭建。管理界面使用 HTML 来进行编写。系统设计使用 SSM 框架。通过对用户画像模型的分析，来达到广告的定点投放。

进度安排：12 月-1 月查阅相关资料；1 月-2 月开始系统的设计；2 月-3 月进行系统实现；3 月 -4 月进行毕业论文的编写。

Student Signature:

陆荣志

Date:

2017.12

指导教师意见

Comments from Supervisor:

同意开题。

1.同意开题

2.修改后开题

3.重新开题

1.Approved(√)

2. Approved after Revision ()

3. Disapproved()

Supervisor Signature:

周凡

Date: 2018, 11

表二：毕业论文（设计）过程检查情况记录表
Form 2: Process Check-up Form

指导教师分阶段检查论文的进展情况（要求过程检查记录不少于 3 次）

The supervisor should check up the working process for the thesis (design) and fill up the following check-up log. At least three times of the check-up should be done and kept on the log.

第 1 次检查（First Check-up）：

学生总结

Student Self-summary:

确立了毕业项目的需求，以及分析了所需要的技术。决定使用 ssm 框架进行毕业设计的开发。

指导教师意见

Comments of Supervisor:

论文进度要加快一点。

第 2 次检查 (Second Check-up):

学生总结

Student Self-summary:

由于整个系统的代码编写还没有完成，所以毕业设计论文还有一部分没有撰写。目前完成了系统的功能分析，各个模块的用例设计。

指导教师意见

Comments of Supervisor:

1. 设计说明需要修改。
2. 论文的章节格式要注意遵守。
3. 目前版本可读性有待提供，系统需求和结构设计需要完善。

第 3 次检查 (Third Check-up):

学生总结

Student Self-summary:

根据互联网行业的广告需求，提出了系统功能：广告管理、广告获取、广告定点投放。然后介绍了各个功能的系统流程，以及精准投放算法的设计。并且介绍了系统架构、各个模块的设计、数据库设计、一些缓存算法的介绍。

指导教师意见

Comments of Supervisor:

文章从系统需求分析到系统各个模块的设计，并且根据广告后台系统的需求，进行广告系统的分布式设计。对于用户画像，可以更加深入。文章结构需要优化，前两章主要描述别人的工作，后面集中表达自己的设计。每一章最好有个小结。最后一章应该对全文进行一次总结。

学生签名 (Student Signature):

日期 (Date):

指导教师签名 (Supervisor Signature):

日期 (Date):

总体完成情况

(Overall
Assessment)

指导教师意见 Comments of Supervisor:

良。

1、按计划完成，完成情况优 (Excellent): ()

2、按计划完成，完成情况良 (Good): (√)

3、基本按计划完成，完成情况合格 (Fair): ()

4、完成情况不合格 (Poor): ()

指导教师签名 (Supervisor Signature):

	日期 (Date) :
--	-------------

表三：毕业论文（设计）答辩情况登记表

Form 3: Thesis Defense Performance Form

答辩人 Student Name		专 业 Major	
论文（设计）题目 Thesis（Design） Title			
答辩小组成员 Committee Members			
答辩记录 Records of Defense Performance:			
<div style="text-align: right;">日期（Date）：</div> <div style="text-align: left; margin-top: 80px;">记录人签名（Clerk Signature）：</div>			

学术诚信声明

本人所呈交的毕业论文，是在导师的指导下，独立进行研究工作所取得的成果，所有数据、图片资料均真实可靠。除文中已经注明引用的内容外，本论文不包含任何其他人或集体已经发表或撰写过的作品或成果。对本论文的研究作出重要贡献的个人和集体，均已在文中以明确的方式标明。本毕业论文的知识产权归属于培养单位。本人完全意识到本声明的法律结果由本人承担。

本人签名：

日期：

Statement of Academic Integrity

I hereby acknowledge that the thesis submitted is a product of my own independent research under the supervision of my supervisor, and that all the data, statistics, pictures and materials are reliable and trustworthy, and that all the previous research and sources are appropriately marked in the thesis, and that the intellectual property of the thesis belongs to the school. I am fully aware of the legal effect of this statement.

Student Signature:

Date:

说 明

1. 毕业论文（设计）的写作格式要求请参照《中山大学本科生毕业论文的有关规定》和《中山大学本科生毕业论文（设计）写作与印制规范》。
2. 除完成毕业论文（设计）外，还须填写三份表格：
 - （1）表一 毕业论文（设计）开题报告；
 - （2）表二 毕业论文（设计）过程检查情况记录表；
 - （3）表三 毕业论文（设计）答辩情况。
3. 上述表格均可从教务部主页的“下载中心”处下载，如表格篇幅不够，可另附纸。每份毕业论文（设计）定稿装订时应随同附上这三份表格。
4. 封三是毕业论文（设计）成绩评定的主要依据，请认真填写。

Instruction

1. Please refer to '*The Guidelines to Undergraduate Graduation Thesis (Design) at Sun Yat-sen University*' and '*The Writing and Printing Format of Undergraduate Graduation Thesis(Design) at Sun Yat-sen University*' for anything about the thesis format.
2. Three forms should be filled up before the submission of the thesis (design):
 - （1）Form 1: Research Proposal of Graduation Thesis.
 - （2）Form 2: Process Check-up Form.
 - （3）Form 3: Thesis Defense Performance Form.
3. All the above forms could be downloaded on the website of the Office of Education Administration. If there is not enough space in the form, please add extra sheets. Each thesis (design) should be submitted together with the three forms.
4. The form on the inside back cover is the grading sheet. Please fill it up before submission.

摘 要

互联网行业中，广告收入占了企业收入的很大比例。并且，互联网用户群体数量日益增长，广告面向的用户也是一样。如何能够有效率地管理、投放广告，具有重要研究意义。

本毕业设计首先会研究分析目前互联网行业中广告系统的需求，提出设计目标——一个方便管理，可用性高的广告后台管理系统；然后文中将会分析目前市面上常用的技术与系统设计框架，选择适合的技术进行开发；其次，本文将会基于外部用户画像系统的建模数据，设计出一个用于广告精准投放的算法，用于广告定点投放以提高广告效率；最后，作者会对整个系统的服务器架构进行说明，设计出一个可用性强的健壮系统。在文中，对于一些重要的算法，也会详细说明。

根据本文中的各个系统模块以及推荐算法的设计方案，作者实现了一个可用的基于用户画像的广告后台系统。系统功能包括三大部分：广告管理，广告获取以及广告精准投放。广告管理包括广告增加、删除、修改，可由管理员手动进行以及系统定时清理过期广告。系统提供可视化界面供管理员使用广告管理功能。

关键词： 广告管理；广告定点投送；用户画像；分布式系统设计

ABSTRACT

In the Internet industry, advertising revenue accounts for a large proportion of corporate income. Moreover, the number of Internet users is increasing, as is the advertising-oriented users. How to efficiently manage and launch advertisements has important research significance.

The graduation project will firstly study and analyze the current needs of the advertising system in the Internet industry, and propose a design goal—an advertising management system that is easy to manage and has high availability. Then the paper will analyze the technical and system design frameworks that are currently used and choose the appropriate technology for development; Second, this article will model data based on external user profile system, and design an algorithm for accurate placement of ads for targeted ads to increase advertising efficiency; Finally, the author explains the entire system and the server architecture and designs a robust system with strong usability. In the text, some important algorithms will also be described in detail.

According to the design of each system module and recommendation algorithm in this paper, the author implements a usable user portrait-based advertising back office system. The system features include three major parts: advertising management, advertising acquisition, and accurate advertising. Advertising management includes the addition, deletion, and modification of advertisements, which can be performed manually by the administrator and the system regularly cleans up expired advertisements. The system provides a visual interface for administrators to use ad management features.

Keywords: advertising management; accurate advertising; user portraits;
distributed system design

目录

- 第一章 引言..... 1
 - 1.1. 背景和意义..... 1
 - 1.2. 设计的目的和意义..... 2
 - 1.3. 本文工作..... 2
 - 1.4. 论文结构简介..... 2
- 第二章 广告系统综述..... 4
 - 2.1. 国内外研究现状..... 4
 - 2.2. 前端技术介绍..... 5
 - 2.2.1. jQuery 框架..... 5
 - 2.2.2. Ajax 技术..... 6
 - 2.3. 后台系统技术介绍..... 6
 - 2.3.1 MVC 设计思想..... 6
 - 2.3.2. SSM 框架..... 7
 - 2.4. 本章小结..... 8
- 第三章 功能需求分析..... 10
 - 3.1. 系统功能..... 10
 - 3.2. 系统运行流程..... 12
 - 3.2.1. 广告获取..... 12
 - 3.2.2. 广告管理..... 13
 - 3.3. 本章小结..... 14
- 第四章 广告系统模块设计与实现..... 15
 - 4.1. 用户画像..... 15
 - 4.2. 精准投放算法..... 16
 - 4.3. 分布式服务器架构..... 17
 - 4.4. 后台业务系统设计..... 18
 - 4.5. 数据库设计..... 20
 - 4.6. 文件缓存..... 23
 - 4.7. 本章小结..... 24

第五章 总结.....	25
参考文献.....	26
致 谢.....	27

第一章 引言

1.1. 背景和意义

广告系统是互联网产业中一个很重要的产业链。在很多企业中广告系统提供的收入占了很重的一个比例。截至 2016 年 12 月，我国网民规模达 7.31 亿，普及率达到 53.2%^[1]。互联网的覆盖面是越来越广泛，则互联网广告的市场也随即扩大。2017 年我国互联网广告市场规模一直处于高速增长的态势，预计 2018 年整体规模有望突破 4000 亿元。在这么庞大的一个市场下，设计一个有效的广告管理系统十分有意义。并且，如何做到高效率地投放广告成为企业的一个重要目标。广告的投放如果随意、泛滥，那么会影响用户体验，严重会导致用户流失。而广告精准定位，有以下好处：提升广告效率，基于大数据精准定向，能够在广告和用户体验中找到较好的平衡；对用户而言，从主动获取广告信息，变为被动需求信息的获取。

本毕业设计来源于数字家庭业务实验中心的广告投放系统设计。数字家庭业务应用是一个注重家庭智慧健康的项目，项目致力于将家庭和互联网科技联系起来，打造一个为家庭服务的数字家庭系统。项目包括一些家庭常用设备的家庭技术支持，如电视机、门口“魔镜”、移动设备（手机）。这些技术支持也就是传统意义上的客户端系统，换言之，这些系统都是可以作为广告投放的载体的^[2]。

互联网广告已经发展成一个至关重要的产业链，其带来的收益对一个互联网产品来说是不可缺少。互联网产品中，盈利的方式整体来说，主要分为下面两种方式：一个是产品应用中提供的付费服务；而另一个就是通过广告投放，从厂商中获得巨额的收益。在一些付费业务几乎不存在的互联网应用中，就必须拥有一个可靠的广告投放管理系统，来维持该应用团队。我所在的项目组也正是属于上述类型。

互联网广告不同于传统广告^[3]。传统的广告包括海报、电视广告这些面向的用户，都是未经过筛选的，而且是未知的用户对象。互联网广告投放和传统广告的最大不同就是，它不仅拥有庞大的用户群体，而且互联网广告面向的用户几乎

是透明的。这个得益于用户的互联网身份大多数都固定的，所以互联网应用中获取到的用户信息比传统的广告投放系统要多得多。根据这些用户信息，所进行的广告投放可以更加精准、有效[4]。

总而言之，一个能够提供广告管理功能，并且容易维护、对使用者友好，并且能够精准定位到目标用户群体的广告系统，是互联网应用中不可缺少的一部分。

1.2. 设计的目的和意义

由上一章节可知，互联网广告的特性是面向的用户群体庞大以及透明度，所以，本系统设计的目标在于设计出一个稳定可用的广告后台系统。系统的主要功能有三个：一个是广告管理功能；第二个是广告自动精准投放功能[5]；还有就是用户客户端获取广告功能。

广告管理功能包括：增加广告、修改广告信息、删除广告、查询广告内容以及统计广告信息。广告管理功能由管理员进行，管理员的专业知识背景要求不高，为了友好性，广告管理系统提供了容易管理的界面功能。

1.3. 本文工作

本文根据分析广告系统的需求，设计精准投放算法、以及设计主服务器各模块和整个系统的分布式架构。

广告系统的实现方面，使用 Apache Tomcat 搭建 Web 服务器，选择使用 Spring + SpringMVC + Mybatis 作为开发框架。数据库使用 MySQL 与 Mybatis 集成作为数据持久层。前端界面采用基于 jQuery 的 AmazeUI 框架进行开发。前端提供广告管理界面。对于精准投放的大数据处理，使用 Hadoop + Spark 的分布式系统进行开发。精准投放是基于其他系统生成的用户画像来进行的。

1.4. 论文结构简介

本文章分为以下几个模块来进行阐述说明本人的广告后台系统：系统设计需求分析，系统各模块介绍与设计，系统效果展示，技术难点与解决方案。

1) 问题综述。分析国内外广告投放系统的研究现状，以及分析现有的主流

技术与框架，选择特定技术进行开发。

- 2) 系统需求分析。根据互联网行业的发展前景，分析广告后台系统的需求，得到系统设计的最终目的。
- 3) 系统各模块介绍。这一模块将会展示系统各个功能的实现方案、UML 时序图，以及整个大系统的 UML 用例图、类图。介绍整个系统是怎样运行的，数据流动等。
- 4) 效果展示。效果展示从两个角度进行：一个是客户端的角度，在客户端上能够正确获取到目标广告；一个是管理员的角度，管理员能够进行广告管理并且得到正确的反馈；
- 5) 总结与展望。总结遇到的困难以及解决方案。列出在编码或者设计的时候遇到的难点，以及自己解决的思路 and 方案。总结系统的优缺点，提出自己的不足，以及未来进一步扩展的展望。

第二章 广告系统综述

本章节将会介绍国内外对于广告精准投放的研究进展,以及 Web 后台系统相关技术的介绍。

2.1. 国内外研究现状

国外对精准广告投放研究有成就的有以下几位。2010 年 Pak Alexander N 提出一种维基百科与语境广告匹配的方法,提高了广告的投放精度。该研究是一种新的机遇维基百科的文章作为“参考点”的广告选择。2011 年, Hof, Robert D 提出了一种机遇社交网站 Facebook 的社交图谱的数据,挖掘出网络用户的兴趣,并根据其推荐相应的广告,完成精准投放。由于 Facebook 的大数据支持,对其进行数据挖掘,能够将广告准确定位到相对应的兴趣和位置的用户。2013 年,台湾的陈艳秋和谢慧清提出的采用神经网络算法改进了 O2O (Online To Offline) 商务模式中广告投放的精准度。O2O 指的是指将线下的商务机会与互联网结合,让互联网成为线下交易的平台。把广告精准投放给目标用户之后,才会有更多的用户在网上支付购买商品,并且在实际的线下得到对应的商品。

国内的研究有以下几位。2013 年,郭新宇、张荣提出一种基于用户搜索行为的潜在语义的用户分割方法。2014 年,张晓阳提出基于 cookie 的精准广告投放技术,根据用户的 cookie 行为信息进行精准广告投放。投放精度提高的同时,必须对用户隐私进行保护[6]。

目前,有几个主流的广告引擎,包括 Google AdMob, eBay, Ikman.lk 等[7]。Google AdMob 是广受欢迎的广告推送引擎之一。这是一个专为移动应用程序设计的广告平台。eBay 是通过互联网进行销售的电子商务公司。它主要使用 Web 应用程序。广告商也可以向 eBay 提供广告。eBay 使用机器学习机制来识别用户的搜索查询并根据之前的搜索和购买历史推送广告。Ikman.lk 是流行的斯里兰卡移动广告网络应用程序之一。用户可以搜索广告,在网站上自由发布广告。

2.2. 前端技术介绍

广告管理系统前端的职责是建立一个简洁、友好、扁平化的用户交互界面。并且负责一些简单的功能：数据校验如输入检验；解析后端数据并显示。

2.2.1. jQuery 框架

广告管理界面采用网页 HTML 的形式来进行开发。而现有的 WEB 前端技术中，使用比较广泛的是 jQuery。

jQuery 是一个轻量级的 JavaScript 的函数库，整体大小只有 30KB 左右，并且 Min 版经过 Gzip 压缩之后，大小只有 18KB。对于 WEB 网页来说是十分优秀的。并且 jQuery 作为一个 2006 年 John Resig 开创的一个开源项目，至今为止，由于越来越多的开发者的假如，jQuery 的功能已经变得身份强大了。可以用很少的代码，完成很多复杂而困难的功能。jQuery 目前集成了 JavaScript、CSS、DOM 以及 Ajax，用起来可以说是方便至极。



图 2.1 基于 Amaze UI 的广告管理前端页面

本广告系统前端页面使用的是基于 jQuery 的 Amaze UI。Amaze UI 是中国首个开源的 HTML5 跨屏前端框架。选择使用 Amaze UI 作为前端开发框架是基于以下优点的：

- 1) 开源。这也是最重要的一点。作为学生并没有经费来购买其他前端框架；
- 2) 轻量级，高性能。和 jQuery 一样，整个 Amaze UI 的函数库文件大小也

是非常可观的；

3) 组件丰富，模块化。Amaze UI 含近 20 个 CSS 组件、20 余 JS 组件，更有多个包含不同主题的 Web 组件，可快速构建界面出色、体验优秀的跨屏页面，大幅提升开发效率。

4) 扁平化设计。简单来说，就是好看。

2.2.2. Ajax 技术

Ajax 的全称为 “Asynchronous Javascript And XML”，则异步的 JavaScript 和 XML，是一种用于创建快速动态网页的技术。

Ajax 技术始于 1998 年，到现在几乎每个 WEB 网页都会使用到该技术。传统的网页如果需要更新内容，必须重新加载整个页面。而 Ajax 技术能够很好的解决这个问题，无需刷新即可动态更新页面。

广告系统前端采用 Ajax 异步请求的方式，与后端进行数据的交互。

2.3. 后台系统技术介绍

2.3.1 MVC 设计思想

MVC 的全称为 Model View Controller，则模型——视图——控制器。这是一种将逻辑、数据、界面显示分离的方法组织代码模式。

View(视图)是应用中负责界面显示的部分，也就是用户能够看得到的页面；Controller(控制器)是应用中负责处理用户请求，与用户交互的部分；Model(模型)是应用程序中用于处理应用程序数据逻辑的部分。

View 层是根据 Model 层来进行创建的；而 Controller 则负责从 View 层读取数据，然后把数据发送给 Model 层；Model 层负责数据的持久化——则数据库的存取。

采用 MVC 设计模式的原因主要是想要把 V 层和 M、V 层分离开来，则前后端分离。这样做的好处是，代码能够模块化，前端的功能只需修改 View 层，而不影响到后端的代码。同理，如果后台来了新需求也是一样，不需要影响到 View

层的架构。MVC 设计模式能够让系统可拓展性更强, 更加容易维护和更加易读。

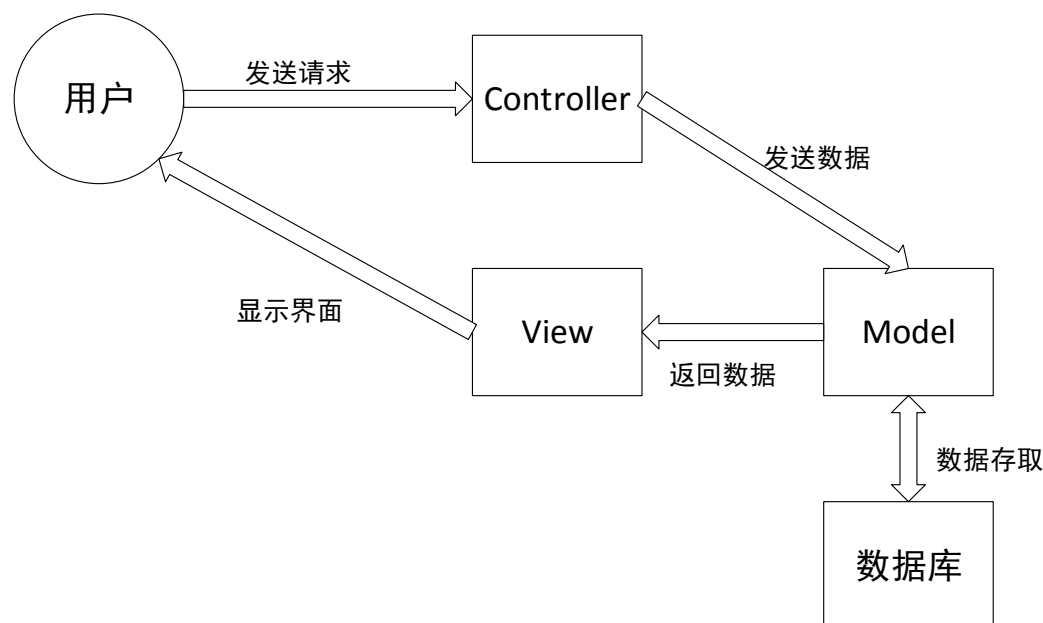


图 2.2 MVC 设计模式

2.3.2. SSM 框架

SSM 框架, 是 Spring + Spring MVC + MyBatis 的缩写。这个是目前比较主流的 Java EE 企业级框架, 也是比较主流的 Web 应用框架。

Spring 是一个 2003 年就开始发行的 Java 开源框架, 其主要的的作用在于使用 JavaBean 来进行开发。JavaBean 可以理解成一个实体类。Spring 带来的技术好处有两个: 控制反转和面向切片编程。面向切片编程是一种编程思想, 面向对象编程也是一种编程思想。本项目并没有很多用到这个功能。而用的比较多的是控制反转。控制反转允许我们交给 Spring 去控制一个对象的实例化和销毁。应用的比较多的是: Spring 将前端传过来的 Json 数据, 解析转换成特定的对象, 极大简化了编程。

Spring MVC 则是一个基于 MVC 思想的 WEB 框架。Spring MVC 很好的把 M 层、V 层、C 层从 WEB 项目中抽离出来。

MyBatis 一个基于 Java 的 MySQL 数据库持久层框架。MyBatis 几乎消除了所有 JDBC (Java DataBase Connectivity, java 数据库连接) 的代码以及参数的手工设置以及结果集的检索。在 MyBatis 中使用简单的 XML 可以将数据库查询结

果映射为 Java 对象返回给调用者。

整体的基于 SSM 的 WEB 框架如图 2.4 所示。

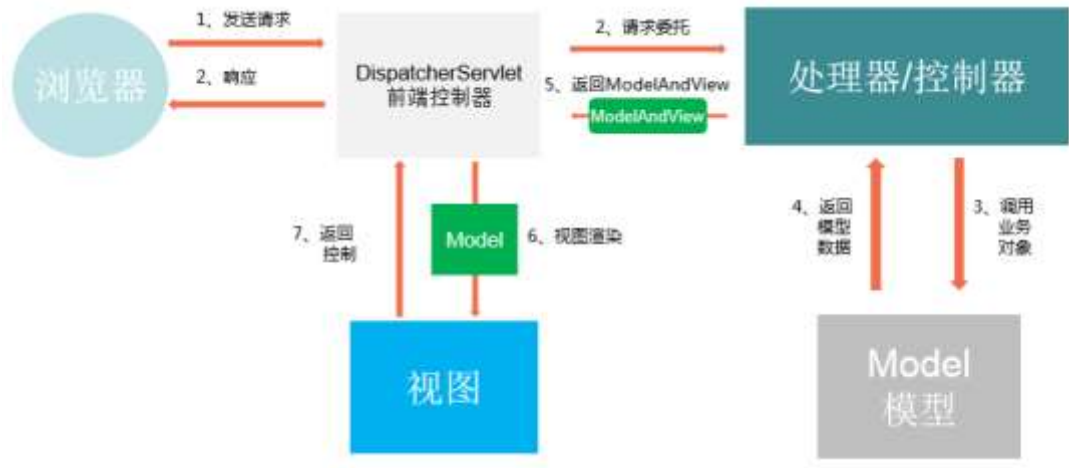


图 2.3 Spring MVC 框架

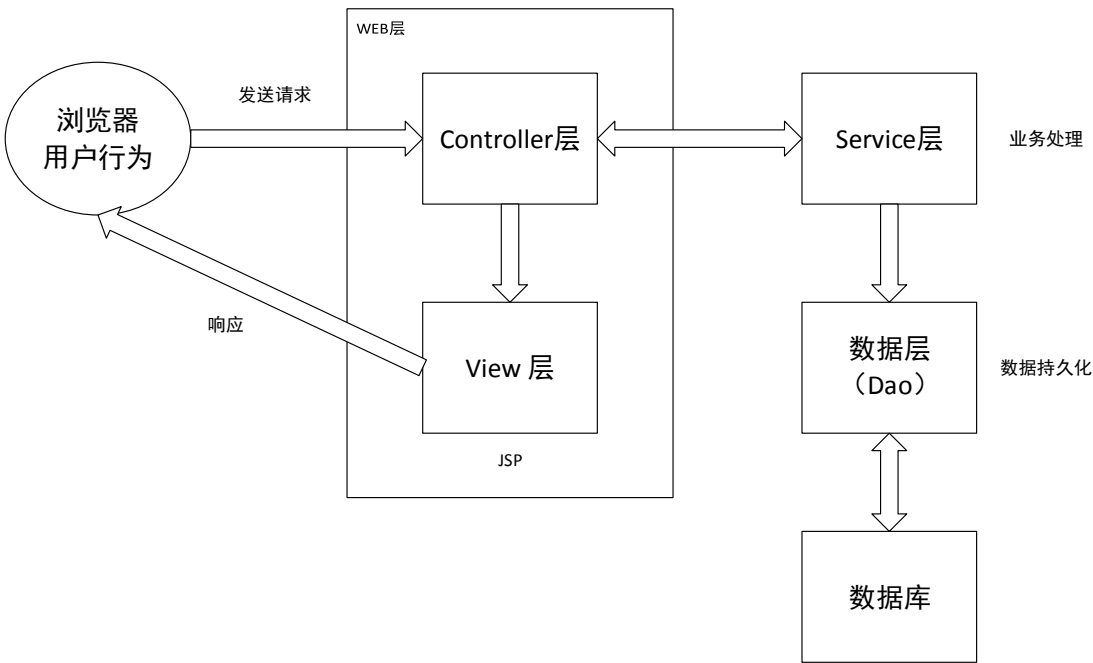


图 2.4 基于 WEB 的 SSM 框架

2.4. 本章小结

本章主要介绍了国内外对于广告投放系统的研究现状，目前研究方向都是首先从用户行为中挖掘出用户数据，建立用户模型，然后根据不同的用户模型进行广告投放。

广告系统使用的服务器为 WEB 服务器,在本章介绍了 WEB 中的前端技术: jQuery 和 Ajax, 后端设计思想 MVC 以及后端 SSM 框架。

第三章 功能需求分析

本设计的广告管理后台系统的目标需求主要有三个：管理员对于后台中广告的管理、目标客户端对广告的主动获取以及广告的精准投放。

3.1. 系统功能

广告管理功能有广告添加、删除、查找以及修改。进行这一列操作需要管理员的权限，则系统需要提供管理员身份的验证服务。

广告自动精准投放功能是基于用户画像系统进行的。用户画像是一组根据互联网应用收集到的一些用户习惯，如用户常访问的页面等等，来使用这些用户信息进行数据分析，预测出来的用户模型。用户画像系统为广告精准投放提供了一个技术支持。建立出的用户画像准确性能够直接影响到广告投放的准确性。由于该原因，本论文设计出来的广告系统为一个与用户画像系统相关联，但是不依赖于某一个特定的用户画像系统的独立系统。广告系统提供相应接口给用户画像系统，来维护一个用户画像模型。而广告系统则根据这个模型来进行定点投放。广告系统是一个独立的系统，脱离了用户画像系统之后，仍可以进行其他两个功能：广告管理和广告获取。

广告获取功能则是一个由目标用户客户端发起的一个功能。广告系统面向的客户端具有松耦合性。则不限制于客户端的类型。无论是常见的 App，又或者是我所在项目组的数字家庭应用的电视平台、网页平台、“魔镜”系统均可使用该系统。出于这个目的，广告投放设计为目标用户主动发起请求进行获取。

整个系统的功能用例图如图 3.1 所示。能够完整作为一个发布的广告模块包括两个部分：用于展示广告的目标应用的用户客户端和用于后台管理的管理员系统。

用户客户端和后台系统是松耦合的，只需要提供能够展示广告的接口，如一片广告区域来展示图片广告，又或者是展示文字亦可。广告后台系统接口具有通用性。在我所在的实验室的数字家庭应用中，客户端包括手机端 App，电视端应用以及网页端应用，这些不同的客户端是可能展示广告的效果不同，但是其从后

台获取广告的方式是一样的。都可以通过 HTTP 请求从后台服务器获取一定格式广告文件，然后再由客户端负责进行展示。

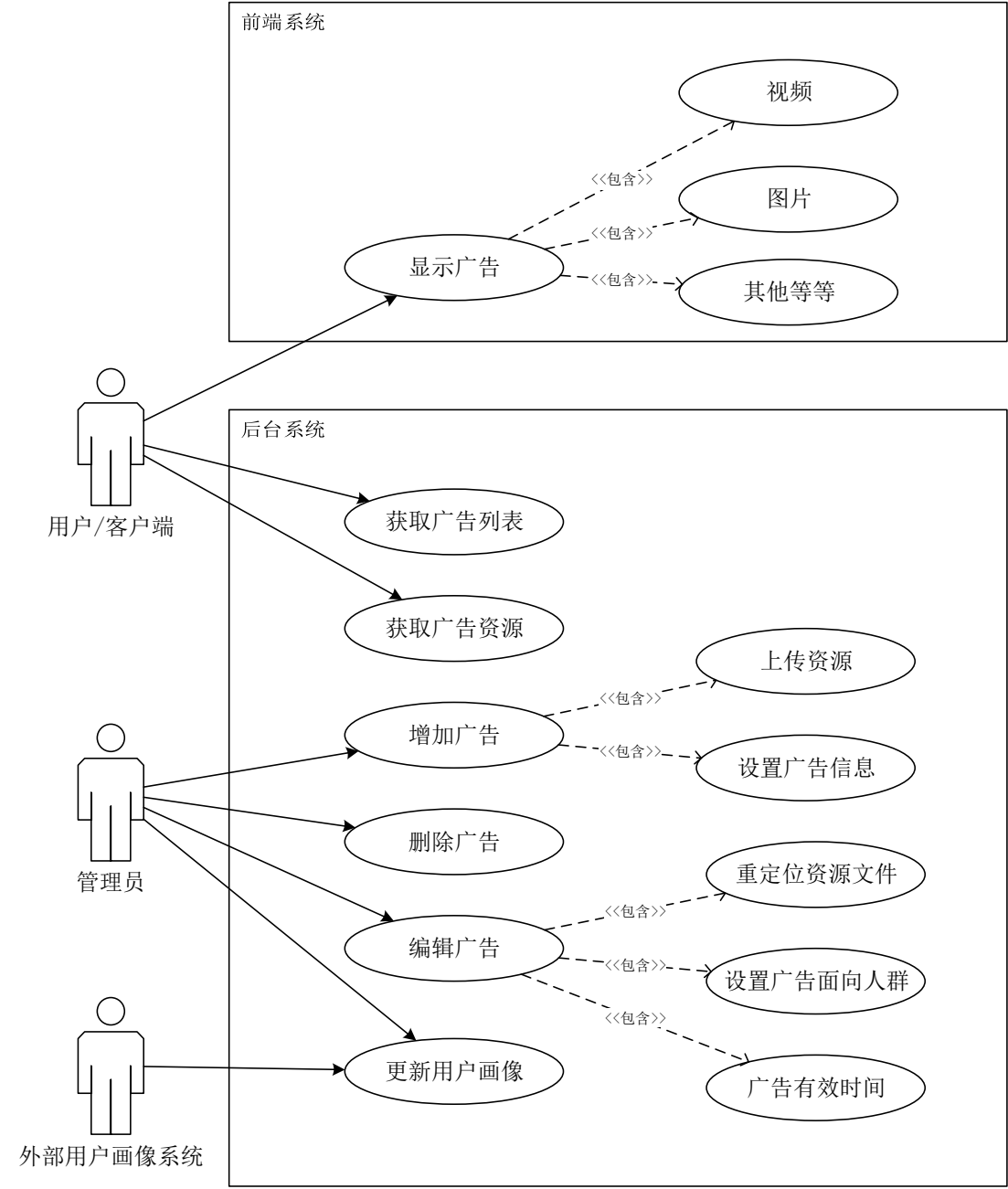


图 3.1 广告系统用例图

用户客户端系统不属于项目内容。用户客户端只需要提供用户的身份标识即可得到用户的投放广告列表，然后再根据投放列表获取相应的广告文件即可。

而另一个模块——后台管理系统目的在于方便管理员高效地进行广告操作：增加、删除、查询以及修改。增加广告包括添加广告信息以及添加广告文件。广告信息包括：广告名字、点击广告后跳转的 URL、广告类型、广告投放时间段、

广告目标人群、广告有效时间。查询广告可以通过广告信息中某一个点来进行模糊查询。修改广告内容则可以修改广告信息和广告文件。

同时，广告后台模块拥有自管理功能，能够定期清理过期的广告，让系统更加人性化。

3.2. 系统运行流程

3.2.1. 广告获取

广告获取由客户端发起。当客户端满足以下条件之一时，发起广告更新请求：

- 广告有效时间已到
- 客户端界面切换到广告界面
- 达到定时获取广告的时间

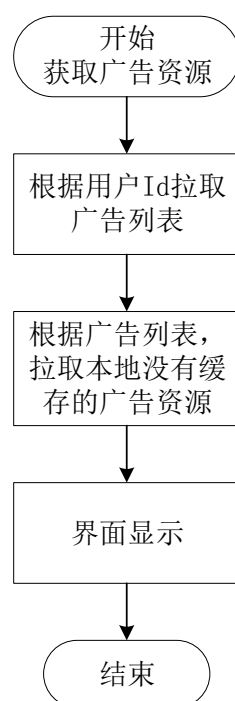


图 3.2 客户端获取广告流程

其中获取广告流程为：先根据用户身份标识给广告服务器发送 HTTP 请求获取广告信息列表。广告信息列表包括了后台系统经过分析之后，给该用户定制的广告的除广告文件之外的所有信息，包括广告投放时间段、广告类型、名字等；

然后根据广告列表，可能客户端本地会做一些处理，来决定自己应该展示什么广告；然后客户端把自己要展示的广告 ID 发送给后台服务器，来下载广告文件，用于前端广告展示。整体系统的时序图如图 3.2 所示。

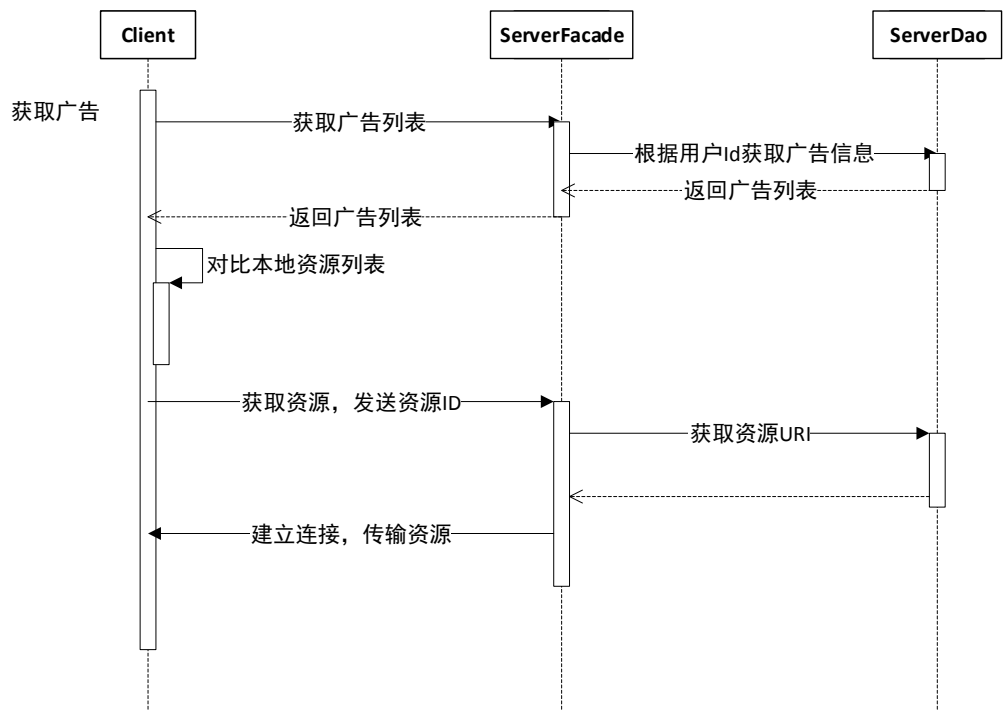


图 3.3 获取广告时序图

3.2.2. 广告管理

该功能属于后台功能。广告的管理有两种：一种是管理员进行添加、删除和修改操作；另一种是系统定期检查更新，剔除过期的广告。广告管理需要进行数据持久化，来保存更新过的广告信息。

广告管理流程见图 3.3 所示。首先由管理员登录到广告后台管理页面，进行广告信息的编写与资源的上传。后台系统得到新修改的广告数据之后，先将广告信息持久化到数据库，然后根据广告的用户标签列表的变化，作出相应修改：如果广告中的用户标签有增加/删除，则在该标签的所有用户的广告列表中增加/删除该广告 ID。然后将修改后的用户和广告的对对应关系更新到数据库。

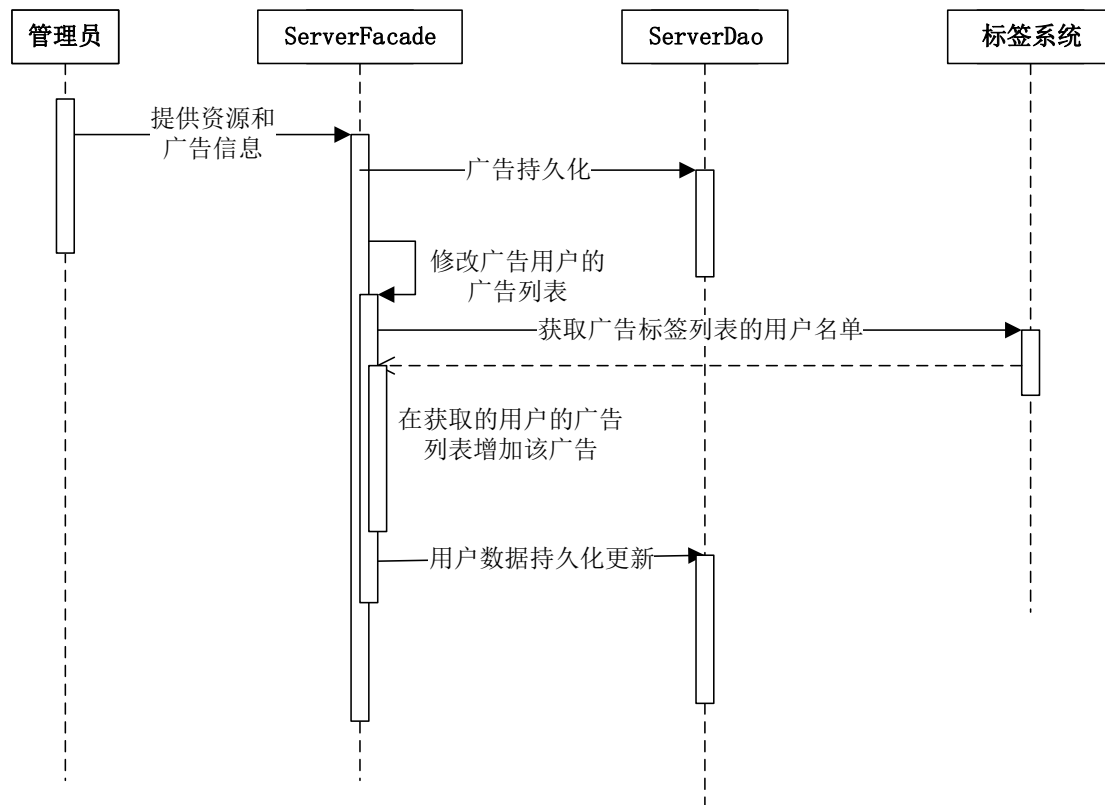


图 3.4 广告管理时序图

3.3. 本章小结

本章节分析了广告系统需要实现的功能，以及系统设计的目标：可用的广告管理界面、用户正确获取广告以及广告定点投放。

本章节还分析了各个功能的调用流程，利用时序图来清晰地展示广告后台系统需要提供哪些接口。介绍了广告获取的流程、广告管理的流程。其中广告获取由客户端发起，这样做的好处是不用保持服务器和客户端的连接，同时也可以实时更新广告列表。广告管理需要多个系统的共同工作，包括数据库系统，用户画像系统以及精准投放计算系统。

第四章 广告系统模块设计与实现

本章节将会介绍基于用户画像的精准投放算法的设计，以及根据第三章中的功能需求，设计出一个松耦合的广告后台系统——使用 MVC 思想，进行各个类的职责划分。并且根据各个服务器的功能，实现一个分布式系统。

4.1. 用户画像

广告的精准投放功能是基于用户画像(User Profile)系统的。用户画像准确度直接影响了广告投放的精准度。广告后台系统不包含用户画像的创建，进行精准投放需要外部用户画像系统的用户模型支持。

用户画像[8]则用户标签，是一种用标签来表示用户的方法，也就是使用标签来为用户建模。标签是一个描述用户的角度，举个例子一个人收入高、坐办公室可以用“白领”标签来表示。

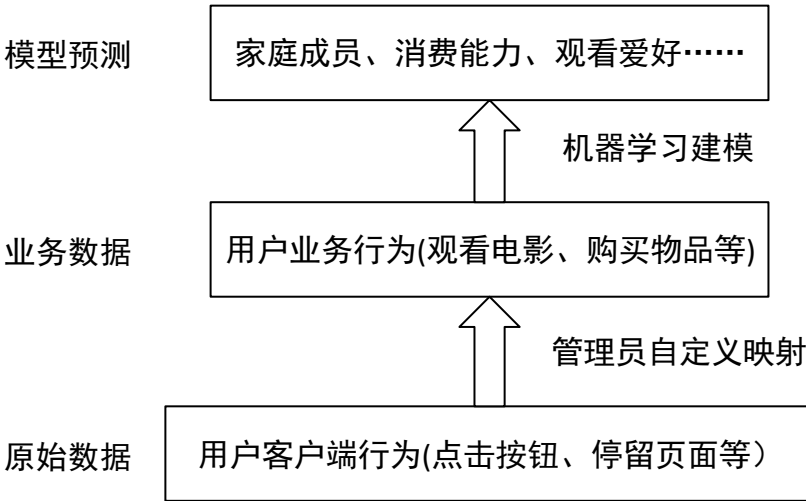


图 4.1 用户画像的建立

用户画像的建立是通过收集用户的日常行为、习惯，然后将收集到的数据放到大数据平台进行分析，建模而得到的。用户画像数据的采集[9]，一般是从很细的行为开始，如常用软件的时间，甚至使用 App 时手指滑动的速度，这些都能成为有贡献的数据。用户画像的生成如图 4.1 所示。

4.2. 精准投放算法

有了用户画像系统的支持，广告系统的定点投放就变得简单了。用户画像将用户表示成如：（学生：0.8，追星：0.5……），这样的数字模型。我们只需要在广告管理中，也为广告添加这样的目标用户范围，然后求出用户与广告的目标用户的吻合度，就可以进行定点投放。前提是广告管理者要对添加的广告熟悉，知道自己的广告的面向群体，这个问题不大。

广告与用户吻合度匹配算法方面。最简单的就是直接判断用户的画像是不是广告目标用户范围的子集。如果是，则把广告加入到用户的广告列表。这是我一开始使用的算法，虽然对于精准投放有一定的影响，但不能作为一个好方法。

其存在的缺点有：第一，不能完全利用好用户画像标签，只存在于符合目标用户和不符合这两个选项，而实际上用户标签是有权重的，上述算法并没有用到用户画像的权重；第二个缺点是，对于一些很“大众”的，用户标签类型不突出的用户，广告的投放的精准度不够。直接匹配的算法，无法更好的分辨出用户之间的细微差别；第三个缺点是，用户广告列表没有一个很好的数量限制，会出现某些用户广告列表上百上千的情况。

为了解决以上缺点，对算法进行改进。上述缺点的原因一个是在于用户画像的权重没有得到很好的利用，另一个是在于用户的广告列表没有一个明确的限制。

为了利用好用户画像，可以对广告的目标用户标签与用户的用户标签通过余弦相似度求出其相似度权重，然后将权重作为精准投放的标准。余弦相似度是一个广泛用于推荐系统的算法。余弦相似度反映的是两个对象之间的吻合度。

$$\cos \langle A, B \rangle = \frac{\sum(A_i * B_i)}{\sqrt{\sum A_i^2} * \sqrt{\sum B_i^2}} \quad (4.1)$$

而对于用户广告列表的限制，可以对相似度权重进行排序，优先选择权重较大的广告进行投放。用户在拉取广告的时候，后台先对广告进行截取，根据策略不同对广告列表进行筛选，如只选择与用户相似度大于 0.5 的前 20 个广告作为投放对象等等。

4.3. 分布式服务器架构

除了外部提供的用户画像系统之外，整个广告系统的服务器分为 4 个部分。分别是广告管理服务器、大数据处理服务器、通信服务器以及数据库服务器。

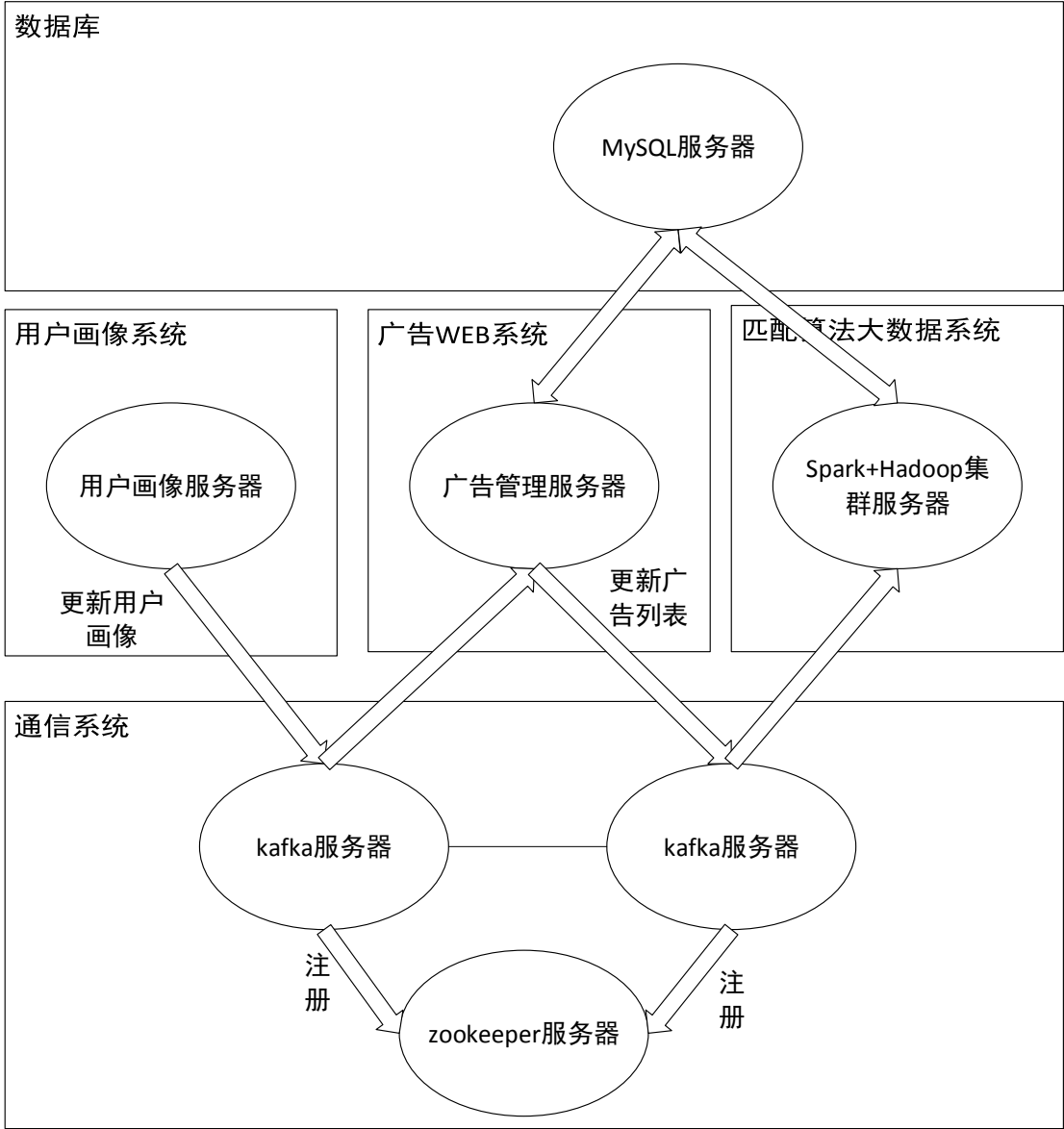


图 4.2 服务器架构

用户画像服务器为外部服务器，主要提供用户标签的更新维护。用户画像服务器不属于广告系统服务器。但是广告系统的精准投放依赖用户画像服务器。用户画像需要对用户进行建模，然后把建立的用户画像模型实时地通知广告系统服务器。所以广告系统提供用户画像的增加、删除、修改接口。并且广告系统有自己的用户画像数据库，来进行自己的精准投放。而广告系统的用户标签的持久化

和维护，需要用户画像服务器调用广告系统的接口来进行。

广告管理服务器为 WEB 服务器，也是主要的业务服务器。负责管理员的广告操作（增加、删除、修改、查询），以及与广告用户的交互，用户广告获取。并且负责后台功能中的广告信息持久化，用户标签持久化，用户广告相似度更新请求的发起。其他服务器都是为了负载均衡的需求才搭建的，所有服务器都是为主服务器而服务。

大数据服务器负责处理广告相似度的计算。这是使用 Spark+Hadoop 搭建的分布式服务器。大数据服务器收到更新广告相似度请求之后，会从数据库服务器中获取数据，进行计算后，把结果再写到数据库服务器。

数据库服务器则负责数据的持久化，作为所有服务器的数据来源。

通信服务器负责服务器之间的通信。kafka 是一个支持分布式的消息队列组件。服务器之间通信采用消息队列方法，对比与采用 HTTP 来说，更加安全可靠，并且响应速度快。HTTP 请求伴随着外网攻击的风险。zookeeper 服务器则用于 kafka 服务的注册。

4.4. 后台业务系统设计

整个广告系统的逻辑集中在主服务器——广告 WEB 服务器中。为了让设计出来的广告后台系统可读性强，并且功能方面容易扩展，能够满足后续的其他功能需求，代码设计采用面向对象的思想。整体框架采用 MVC 框架，把整个大的系统按照功能分为一个个模块，做到各个模块各尽其责，各个模块只负责单一任务。解耦的好处在于，当遇到 bug 或者新需求的时候，能够快速定位到问题所在。某一模块的改变，不会影响到其他模块。人的大脑和电脑硬件有相似的地方，当程序员看代码的时候，会把代码装载进大脑，系统解耦得越好，则遇到问题的时候大脑需要考虑的地方就越少，效率就越高。

根据 MVC 思想以及 SSM 框架，我将广告后台系统分为四层：View 层，Controller 层、Service 层以及 Dao 层。

View 层负责网页的前端展示。其主要载体为 jsp 文件；Controller 层负责接收用户请求，然后分发到 Service 层；Service 层负责业务逻辑的处理，不同的业务有不同的 Service；而 Dao 层则负责数据的持久化和读取。

整个业务响应流程为：View 层负责界面的显示以及用户的交互，View 把接收到的用户请求交由给 Controller 层进行处理，Controller 根据用户请求区分成不同的业务调用 Service 层（业务层），Service 层则可以调用 Model 层的 Dao 进行数据的持久化，则保存到数据库。然后把处理过后的信息再逆向通过 View 层反馈给用户。

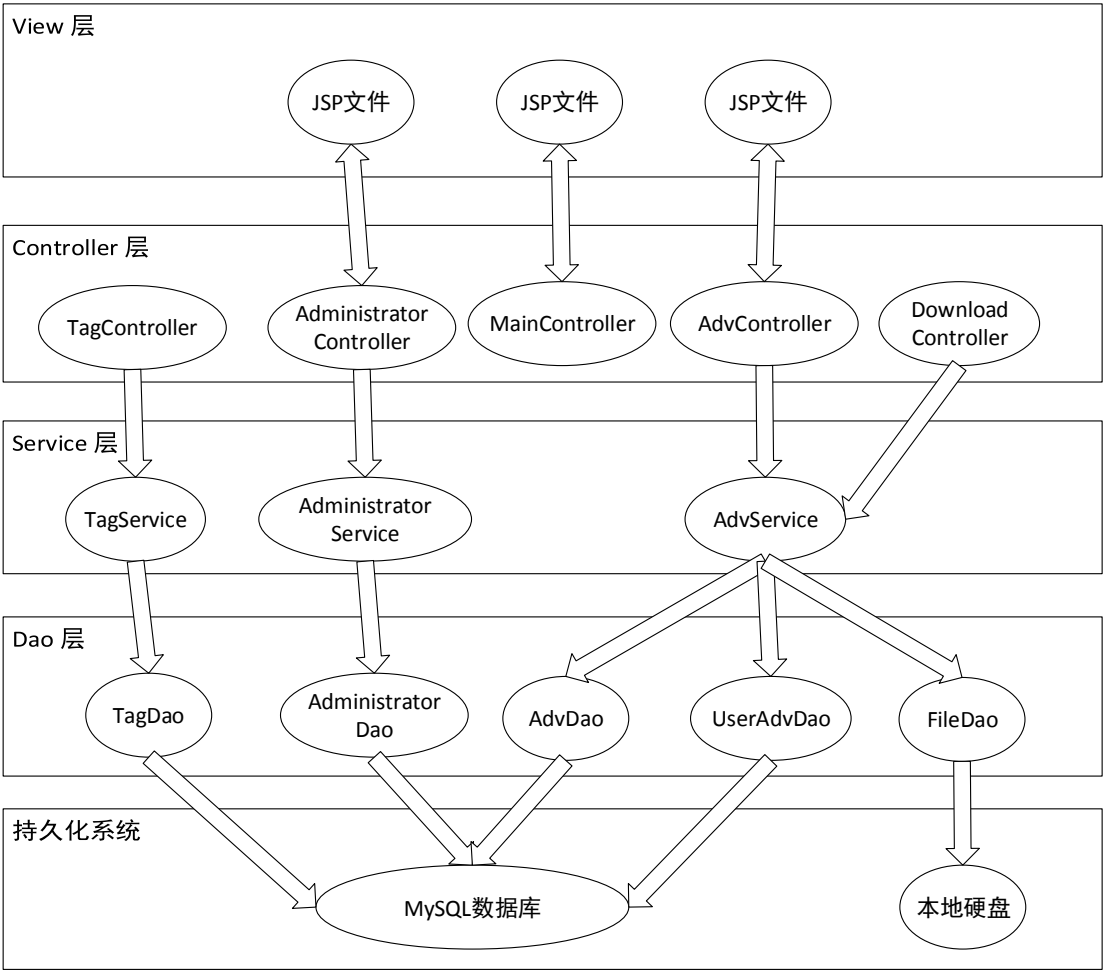


图 4.3 后台系统分层模块

如图 4.3 所示。一个 Controller 可以调用多个 Service，一个 Service 也可以调用多个 Dao。Dao 设计的原则是尽量一个表格一个 Dao，尽量不要存在多个 Dao 同时对一个表格的操作。事实上，这个很难避免。我在设计的数据库的时候，插入数据操作经常会用到联表查询。数据库这样设计的好处在于，一表一个 Dao 容易进行数据缓存管理。数据库的设计将会在后面详细讲解。注意到的是广告文件并不保存在 MySQL 数据库中，因为如果这样做的话数据库的负担会很大，企业一般也不会这样做。

我根据功能的划分，将 Controller 分为 5 个。分别是单纯用于与 View 层界面交互的 MainController；用于管理员登录管理的 AdministratorController；用于外部用户画像系统更新用户画像的 TagController；用于广告信息管理以及广告信息获取的 AdvController；以及用于广告文件下载的 DownController。

根据不同的特定业务，建立了 3 个 Service。分别是用于管理和获取用户标签的 TagService；用于管理员登录密码判断、登录状态维护以及管理员密码修改的 AdministratorService；还有就是负责处理与广告信息相关的 AdvService。其中最为重要的就是 AdvService，在这里面负责了广告的增加、删除、查询、修改，以及负责与广告匹配大数据服务器的通信。

4.5. 数据库设计

数据库表格分为 4 个模块。用于存广告信息的广告表格；用于保存用户画像的用户画像表格；用于保存用户广告列表的用户广告表格；以及用于保存管理员个人信息的表格。

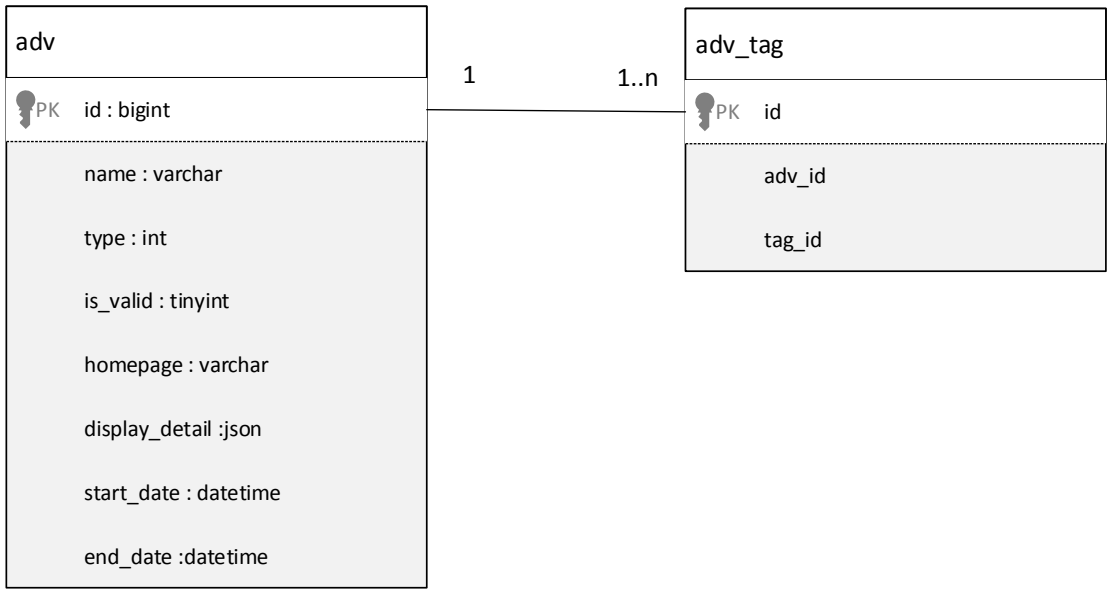


图 4.4 广告表格

广告表格中，目标广告与目标用户标签是一对多的关系，则一个广告拥有多个目标用户标签。表格中 id 字段为主键唯一标识；name 为广告名称；type 为广告的类型，用来表示广告是图片还是视频或者是文字；is_valid 字段表示该广告是否处于有效的状态，如果值为 0 则表示无效，用户将不会获取得到该广告；

homepage 字段为广告跳转的 URL；display_detail 字段用于保存广告的投放时间段，用 json 格式保存便于拓展；start_date 为广告开始的有效时间，end_date 为广告截止的有效时间；tag_id 对应用户画像系统的标签唯一标识。

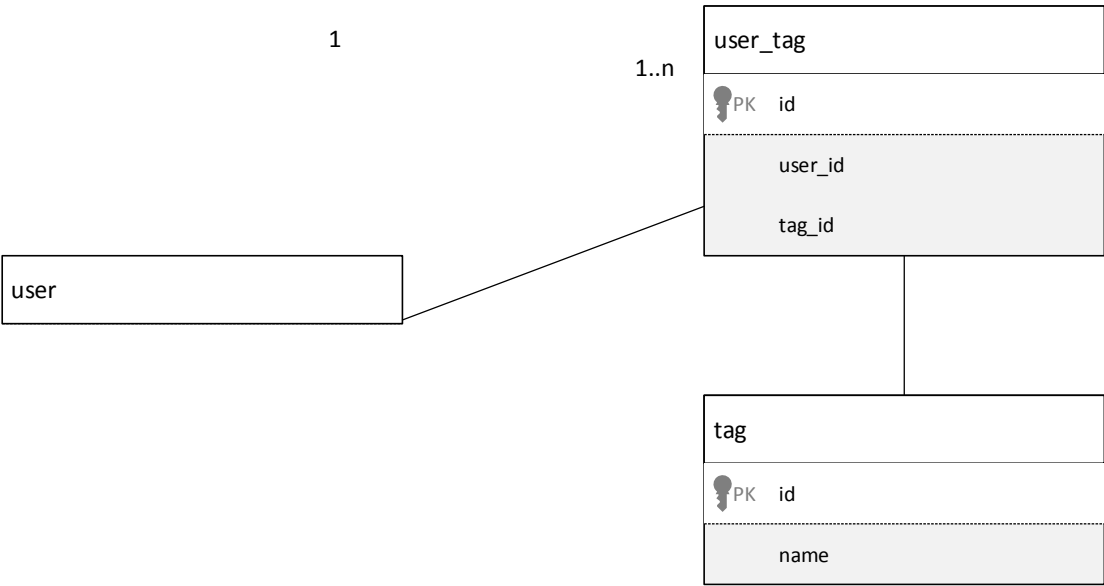


图 4.5 用户画像表格

用户画像表格中，用户和用户画像也是一对多的关系。剩下的管理员表格就是一个 id 和密码字段，没必要单独说明。

值得一提的是，数据库中有一个表格不属于以上三个持久化模块，属于一个业务功能的表格产物。在保存广告文件的时候，我的设计是广告文件的名称是广告 id+文件后缀。所以在保存广告文件的时候，必须获得广告 id。MySQL 中提供了自增 id 字段，以及 last_insert_id() 函数，可以获得最后插入的数据的 id 值。

但是用这个方法来获取新插入广告 id 有两个缺点：

第一，获得广告 id 必须要在广告信息插入到数据库之后，这个和广告后台系统的需求冲突了。广告文件的保存必须要在保存广告信息之前。原因是，如果先保存广告信息到数据库，然后保存广告文件失败，就会出现数据库中的广告信息的错误的，无法得到正确的广告文件，想要解决这种情况唯有数据库的回滚。其开发成本是很高的；

第二个缺点，必须保证插入操作和 last_insert_id() 函数的调用是一组原子操作。在多线程的环境之下，插入操作同时会进行多次。要保证原子性，就要加锁。因为数据库的操作通常是很耗时的，尤其是插入操作。在这里加锁很导致高并发的

时候的线程阻塞。

为了解决这一情况，我写了一个用于获取不重复 id 的模块。整个模块的类图如下图 4.6 所示。IdMgr 类为一个单例[10]。单例可以保证其数据可以在其他类中共享并且可以有效地管理数据。

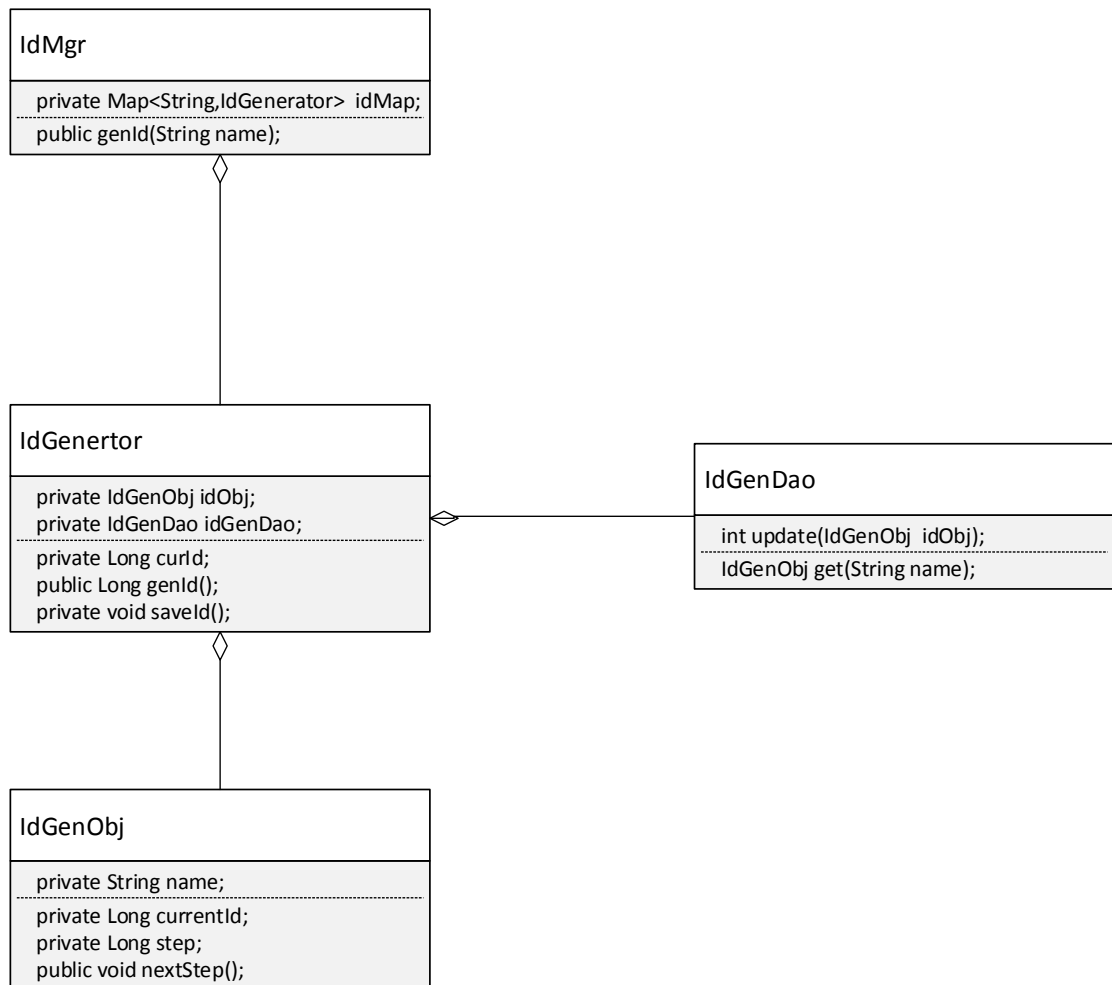


图 4.6 id 生成模块类图

大致的算法是，通过调用 IdMgr.genId(String name) 这个方法来获取 id。IdMgr 为一个单例。整个算法的思路是：IdMgr 创建的时候，会访问一次数据库，读取数据库中 id_generator 表格的所有记录，根据读取的内容生成 IdGenerator，保存到以字段 name 为主键的 Map 中。其中 current_id 字段记录了上一次的未使用的 id，step 字段为每个 current_id 的间隔，则每个 current_id 的值都能整除 step 字段。


id_generator	
 PK	name : varchar
current_id : bigint	
step : bigint	

图 4.7 id_generator 数据库表格

生成 id 时，调用 IdGenerator.genId()方法获取。IdGenerator.genId()的代码如下：

```
public Long genId() {
    synchronized (this.idGenPojo) {
        long maxId = this.idGenPojo.getCurrentId();
        this.curId += 1;
        if (this.curId >= curId) {
            idGenPojo.nextStep();
            UpdateDb();
        }
        return this.curId;
    }
}
```

算法思路为：保存一个 Long curId 的变量，每次调用 genId()方法的时候让其加 1。然后判断 this.curId 是否大于数据库中的 current_id 字段，如果是则让 current_id = current_id + step，然后更新数据库中的 current_id 字段。最后返回 curId。

4.6. 文件缓存

之前提到过，广告文件不保存在数据库中以减轻数据库的压力。而保存在磁盘的文件，在读取的时候会耗费大量的时间。而广告文件的获取是十分频繁的。为了解决上述情况，需要对广告文件进行缓存。

伪代码如下：

File getFile(String filename):

File newFile

IF filename 缓存不存在:

newFile = 读取磁盘文件

把 newFile 加到缓存

Return newFile

在 FileDao 模块中，采用 Google 提供的 Gauva 中 LoadingCache 作为缓存中间件。选用 Gauva 的原因是其组件是线程安全的，并且容易使用和管理。缓存的流程为：读取广告文件的时候，先访问缓存中是否有该文件，如果有则直接从缓存获取，无则从本地磁盘读取文件。有了缓存，只需要在第一次读取的时候访问磁盘，大大提高效率。在保存或更新广告的操作中，需要先清除缓存中的对应的广告文件，然后再保存到本地中，以避免脏数据。

4.7. 本章小结

本章主要介绍了用户画像的概念以及用户画像的大致创建流程，然后根据用户画像设计出一个广告定点投放算法。用户画像模型是定点投放的关键。

根据应用服务的不同，将广告系统分为业务服务器、数据服务器、通信服务器、大数据处理服务器这四个服务器模块。

主要的设计思想应用在业务服务器上面。本章使用 MVC 的设计思想，SSM 的 Web 框架来进行系统设计。根据业务服务器中的功能划分，分为几个独立的模块，来达到解耦的效果。

第五章 总结

论文分析了广告投放系统的需求,提出设计目标:设计出一个拥有广告管理、广告获取、广告定点投放功能的后台系统。后台业务服务器采用 WEB 服务器的形式来进行搭建。采用的技术有 SSM(Spring + Spring MVC + MyBatis)框架, Apache Tomcat 服务器搭建, Kafka 消息队列, Hadoop + Spark 大数据计算平台。

其中广告获取由客户端发起,流程为客户端发送 HTTP 请求获取用户的目标广告列表,获取列表后根据其中信息向服务器再次发送 HTTP 请求下载广告文件用于展示。

广告管理为管理员功能。广告管理提供管理员身份验证,使用拦截器和 Session 验证进行对没有登录的人员的页面拦截。广告管理提供扁平化界面供管理员使用其广告内容管理功能。在广告管理网页平台上,管理员可以添加广告、删除广告、查询广告内容以及修改广告的内容。广告内容包括:广告名字、跳转主页、有效时间范围、投放时间段、投放目标人群、广告文件。管理员对广告内容进行修改更新后,将会由大数据计算平台进行广告精准投放的计算。

广告精准投放是一个基于外部用户画像系统的算法。用户画像则用户模型,是一种用标签和权重来表示描述用户的方法。用户画像由外部系统去维护、更新,广告后台系统提供用户画像模型数据的收集接口。当外部用户画像系统中对于用户的标签与权重有所更新,就会调用广告后台系统的更新用户画像接口。广告后台系统接收到更新请求后,会更新数据库中的用户画像,并且重新计算相关的用户广告列表的权重。

精准投放算法为求出用户标签与广告目标用户标签之间的余弦相似度,然后根据其权重对用户广告进行筛选。由于用户数量和广告数量,余弦相似度计算量比较大,所以使用大数据计算服务器来运行计算。

根据以上后台功能的不同,以及处于负载均衡,系统性能的考虑,将广告后台系统分为四个子系统。一个是负责主要的广告功能的 Web 服务器,如广告管理、广告获取;另一个是负责数据持久化的数据库 MySQL 服务器;另一个是集成 Hadoop + Spark 的大数据计算服务器,用于计算用户广告权重;还有一个是 kafka 通信服务器,负责 Web 服务器与大数据服务器之间的通信。

参考文献

- [1]. 第39次《中国互联网络发展状况统计报告》[J]. 中国经济报告,2017(04):7.
- [2] 栾俊华. 论智能终端广告系统的分析与设计[D]. 北京邮电大学,2012.
- [3] 徐国建. 广告管理系统设计与实现[D]. 电子科技大学, 2009.
- [4] 张建, 孙铭, 段娟. 基于大数据平台的精准广告系统研究与设计[J]. 电脑与信息技术,2015,23(04):47-50.
- [5] 逢山舒.基于大数据时代下的精准广告应用研究[J/OL].现代营销(下旬刊),2018(01):66[2018-04-19].
- [6] 高杰. 基于用户行为的精准广告投放研究[D].武汉工程大学,2016:3-5
- [7] Hansi De Silva, Poorna Jayasinghe; Ashen Perera, Sithira Pramudith; Dharshana Kasthurirathna. Social media based personalized advertisement engine[D]. 11th International Conference on Software, Knowledge, Information Management and Applications (SKIMA), 2017:2-3
- [8] 余孟杰. 产品研发中用户画像的数据建模——从具象到抽象[J]. 设计艺术研究,2014,4(06):60-64.
- [9] Gerrit Kasper, Diego de Siqueira Braga, Denis Mayr Lima Martins, Bernd Hellingrath. 2017 IEEE Latin American Conference on Computational Intelligence (LA-CCI):2017:1-6
- [10]. Bloch J.. Effective Java: Programming Language Guide[M]. Addison Wesley. 2001.

致 谢

感谢周凡老师对我本科毕业论文的多次指导，感谢陈湘萍老师对我的开题报告、论文内容的建议和指导，感谢郑贵锋老师对我的项目的方向指导、技术指导以及进度安排。

毕业论文（设计）成绩评定记录

Grading Sheet of the Graduation Thesis (Design)

指导教师评语
Comments of Supervisor:

成绩评定
Grade:

指导教师签名
Supervisor Signature :

Date:

答辩小组或专业负责人意见
Comments of the Defense Committee:

成绩评定
Grade:

签名:
Signatures of Committee Members

Date:

院系负责人意见
Comments of the Academic Chief of School:

成绩评定
Grade:

签名

院系盖章

Signature:	Stamp:	Date: