第十五届 “挑战杯”全国大学生课外学术科技作品竞赛

参赛作品

广州白云机场客流量大数据分析系统
“A-guardian”

二零一七年三月

摘 要

随着新一代信息技术的迅速发展，数据已经深入到当前每一个行业和业务职能领域，成为一种重要的自然资源和生产元素，我们已进入了数据量大，类型丰富，潜在价值高的大数据时代，在这个时代中，亟待人们对这些数据加以合理、高效、充分的利用，使之能够给人们的生活工作带来更大的效益和价值。

本作品 A—guardian 正是基于这样的背景，基于机器学习技术和 spark 大数据平台，应用了 GBRT、OPTICS 点排序聚类等智能算法，实现了对白云机场人流量的实时监控、预测，使航站楼内值机柜台、登机口、广告位等设施安排更为合理，能更高效地调度资源与人员，减少机场费用的同时也保障了机场安全，在功能升级时会将客流数据推送给周围运力单位，实现机场、旅客和运力单位的多赢。同时，本作品基于服务器端，通过网页呈现给用户，只要能浏览网页都能使用我们作品提供的功能，极大的消除了操作平台的差异性。值得一提的是，本作品中使用了优化的 OPTICS 算法，极大的提高了系统分析数据的效率，**已经申请多项专利**。

在创新上，选取多个不同特征、使用不同的算法进行分模块建模，最后将这些子模块加权融合得到本作品的核心系统，本系统能以 10 分钟为间隔，连续动态预测未来 10 分钟、20 分钟直到未来两天的人流分布，实践证明：**在阿里大数据云平台运行后的结果排名为 Top1.96%，准确度高，效果好。**除此之外，本作品还可针对异常的实时或预测的人流分布情况给予不同的、合理化的建议，并将监控和优化前后的情况通过可视化清晰明了的展现在了用户眼前。同时，本系统可在全国机场做推广和复制，通过相应的修改与扩展后适用范围更广，对大多数公共区域都可使用，有利于大数据相关技术与产业的结合。

关键词：数据挖掘　机械学习　实时预测　特征提取　分模块训练

专家推荐意见摘录

王国胤 重庆邮电大学大学计算机科学与技术学院院长 教授 博士生导师

该作品将多种传统机器学习算法融合，并加入新颖的聚类方法，能够对数据进行实时预测，对机场人员调度，设施安放有实际意义。

胡峰 重庆邮电大学大学计算机科学与技术学院 教授

该作品应用具有自主产权的数据聚类方法，数据挖掘技术，能提供有效的模型用于预测数据发展前景好。

王进 重庆邮电大学大学计算机科学与技术学院 教授

该作品能做到数据实时预测和可视化，有良好的市场前景和社会应用价值。

主要创新点

创新点一：针对机场客流分布情况的预测和可视化

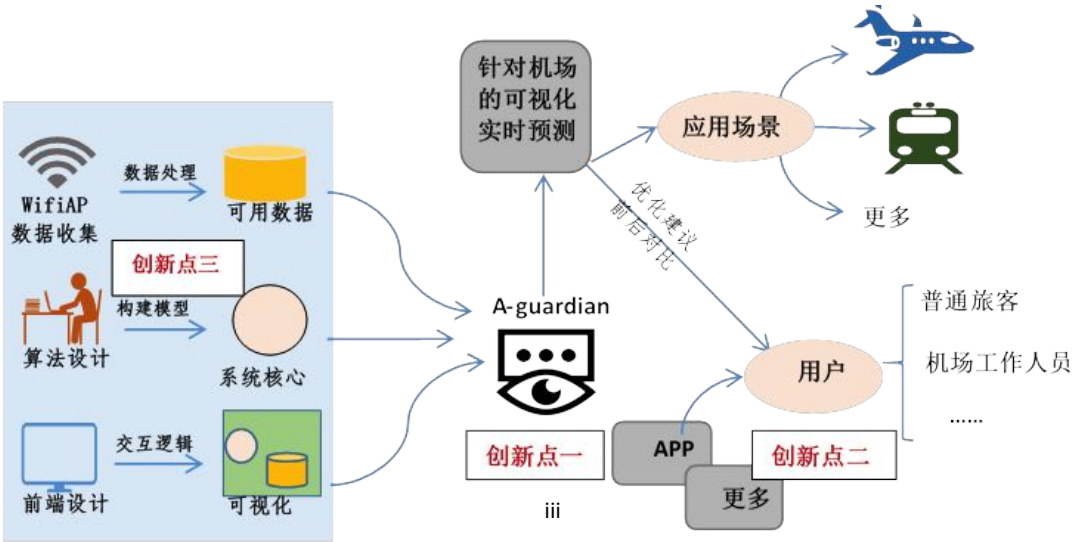
本系统创新地将大数据预测应用在机场人流分布这一全新的场景中，根据机场的人流分布，提供优化建议，使航站楼内值机柜台、登机口、广告位等安排更为合理，能更高效地调度资源与人员，减少机场费用的同时也保障了机场安全。

创新点二：智能地提供预警、运力调配等解决方案

系统功能上，本作品可对机场人流分布情况进行实时监控，以 10 分钟为间隔连续动态且快速地预测人流分布，也可让用户自由设置时间间隔。特别的，本作品设计相应的人流分布预测和预警功能对相关的机场区域做出合理的人员设备调整，同时将人流信息推送给机场周围运力设施，提醒相关客运单位提供旅客接送服务，并将优化前后的情况清晰明了地展现在用户眼前，实现机场、用户和周围运力单位的多赢。

创新点三：Spark 平台下的数据处理和模型构造方法

数据处理上，我们创新地提出了 Spark 大数据平台的 OPTICS 点排序聚类方法检测离群点以及三支决策不平衡数据过采样方法来进行数据预处理；构建特征时，我们多次挖掘桥梁特征将看似无关联的特征联系起来，设置成新特征；建立模型时，构造子模型按权融合，使预测精度更加准确。



目 录

第1章 作品概述

1.1 作品研究背景

在这个科技迅速发展的时代，交通日益发达且人们生活质量逐渐提高，飞机成为了大多数人，尤其是旅游人群的出行方式。机场有巨大的旅客吞吐量，与其相对应的是巨大的服务压力。为了有效利用机场资源，提升生产运营的效率，机场内需要不断提升运行效率的资源有航站楼内的各类灯光电梯设施设备、值机柜台、商铺、广告位、安检通道、登机口，航站楼外的停机位、廊桥、车辆（摆渡车、清洁车、物流车、能源车）。而要想提升这些资源的利用率首先需要知道未来一段时间将会有多少旅客或航班会使用这些资源，其次需要精准的调度系统来调配这些资源和安排服务人员，才能够帮助机场提升资源利用效率，保障机场安全与服务提升。所以，借用大数据技术推动机场业务快速发展是本次作品的创新点。

利用大数据处理技术，本作品以前一周机场旅客分流量为基础，可以预测未来 2 天的机场人流量，以及人流的的大致分布区域。使机场能够提前将对应的配套设施安装到相应位置，将机场的服务人员的工作效率达到最大化，并减少机场旅客的安全隐患，且以较高的服务质量吸引更多的航空公司的合作，及跟随旅客而产生的商业企业的加盟。

1.1.1 机场现状

2012 年国务院提出对民航发展的若干意见指出，当前民航业发展中不平衡、不协调的问题仍较为突出，空域资源配置不合理、基础设施发展较慢、专业人才不足、企业竞争力不强、管理体制有待理顺等制约了民航业的可持续发展。

首先，民航机场作为航空运输的基础设施，是综合交通运输体系的重要组成部分。人们对航空领域的需求量增加，使得各地都兴建机场，直接导致机场行业的竞争越来越激烈。机场服务质量和服务管理水平作为衡量机场优劣的两大基本要素，在客户对机场要求日益提高的今天，对机场的发展起到决定性的作用。如

何不断提高服务质量和服务管理水平，满足旅客日益增长的需求，提升机场企业的核心竞争力，已经是民航机场广泛关注的重要课题。

因此对机场服务质量管理的研究具有重要的现实意义，其具体体现为以下三点：

（1） 由于服务具有求别实体产品的无形性、不可存储型、人员参与性等特点，所以服务质量的管理不能参照实体经济，而要根据现状自成一套管理体系。

（2） 机场服务面向对象是各地旅客，旅客所看重的是机场服务过程中以服务为核心的附加价值，所以无论是从人数，还是国籍，重要的是使旅客有宾至如归的感觉，才能由此带动机场一系列的附加运营产业

（3） 开展机场服务质量管理工作，有效地对服务提供单位进行强有力的约束和规范，有助于提升以机场为核心的企业形象和企业文化。

所以从机场管理角度出发，如何制定有效的管理方法，以提高服务质量，优化机场设施？众所周知，由于机场的特殊性，机场最大的隐患就是人流量的问题。庞大的人流量，影响了机场各种配套设施的使用。从旅客进入机场的那一刻起，办理登记手续，托运行李，安检，候机，甚至领取行李，转其他交通工具，旅客随时都处在排队状态。其直接导致了大量旅客滞留在机场范围，占用大部分的机场设施。由于员工和设施有限，机场不能够为所有客户提供其所能及的服务，服务质量降低，其直接后果是机场的口碑下降，竞争力下降。

其次，由于机场环境的特殊性，考虑到机场建设面积、居民生活环境等一系列因素，我国机场大多建设在远离市区、居民区的地方，所以，“机场周围交通不便”也是近年来，旅客在航空行业反映的较为突出的问题之一。

1.1.2 机场定位

机场是民用航空和整个社会经济的结合点，机场也是一个地区的公共服务设施。因此，机场既带有营利性质，也带有为地区公共服务的社会事业性质。

1.1.3 机场服务对象

【旅客】由于出行需要必须经由机场，在机场完成候机、下机或转站的人群，

有机场提供相应的问询服务、行李托运、安全检查、候机、旅游咨询、餐饮购物等，可以满足其绝大部分需求的服务；

【航空公司】有机场提供运行场地的飞机拥有企业或公司，其所属的飞机包括机组人员，有机场提供相应的跑道，引导等服务；

【与机场相连的其他交通运行设施】包括相应的轻轨、地铁、公交车站、出租车站等。

1.1.4 机场服务范围

按照服务对象，机场武器提供以下服务：

（1） 机场为旅客提供的服务：

核心服务：办理登机、问询服务、行李托运、安全检查等

便利服务：路线咨询、旅游咨询、餐饮购物、休闲娱乐服务等

（2） 机场为航空公司提供的服务：

提供飞机维护、停机坪、起降服务、跑到引导、装卸服务机及航空人员休息等

（3） 机场为其他运营设施提供的服务：

为相邻交通设施提供特殊通道，维护旅客候车秩序等

1.1.5 机场服务漏洞

【安检】由于机场的特殊性，安检强度远远高于其他交通设备，由此导致正常安检时间较机场预计的时间长。其中有异常情况时，对特定旅客的特殊检查仍由原安检人员检查，导致更多旅客排队等候。同时，机场每一天只开放几个安检口使用，导致同一安检口，旅客滞留量增多。

【登机口安排】登机口安排不够合理。由于不同飞机在同一机场的载客量不同，不合理的登机口安排，使很多旅客在安检延迟的情况下还需走很远的路。旅客为了能正常登机，常会提前到达机场，导致一时间段内，机场滞留量上升。

【资源浪费】在机场内并不是每一样设备，每天都会有很多人使用。旅客到机场乘机，基本会以目标登机口最短距离行动。所以很多不在旅客行动路线上的设备会被旅客弃用，白白浪费电力等资源。

【安全】由于机场地方较大，安保人员致其范围广，也造成了紧急情况发生时，不能及时到位。

【相关交通设备】这类交通设备不能及时接送旅客离开，使机场中旅客滞留时间增长，同时也造成了交通拥堵的情况。

针对此背景，我们作品以提前预测将分布人流量为基础，对人流量大的地区提前预警，为机场提前安排适宜的服务设施、调度人员提出合理建议，修改不相应的管理办法。

1.1.6 预计解决情况

经过相应的数据分析，根据系统提出的预警信息，对相关的机场区域做出合理的人员设备调整，将人流信息推送给机场周围运力设施，提醒相关客运单位提供旅客接送服务。

1.1.6.1 预警

【旅客疏散】由于我们是基于大数据平台进行的智能预测，是具有实时性的。所以，当在某个点连接 Wi-Fi 人数逐渐增多至难以控制时，系统就会发出预警信号。相应的，相关管理工作人员就会出动来疏散旅客，避免安全事故的发生，疏散旅客的效率也可以根据该系统的实时性得出。

【机场人员调度】在我们的预测系统中，如果预测到某个地方的人员流动量很大，容易发生危险，不安全的情况时，系统也会发出相应的预警信号。机场就会进行相应的机场人员的调度，从而疏散人群，避免安全隐患。

【设施安放】根据旅客连接 Wi-Fi，从而系统发出实时的人流量分布情况，以及预测系统中预测到的未来几小时的人流量分布，机场方面可以看出哪个点的人流量大，旅客滞留久，就会相应的把一些安全设备，紧急设施放在那些急需的区域。既能够避免安全事故的发生，又能够给旅客提供更优质的服务。

1.1.6.2 运力调配

航空客运作为我国重要的大众运输工具之一，具有舒适、快捷等特性，因而成为了许多民众乐于采用的交通出行方式。随着我国经济的的发展，航空行业近

几十年也取得了长足的发展。针对机场客流表现出的客流惯性进行分析发现，节假日的出行给机场带来巨大客流量的同时也时常伴有旅客滞留、机场拥堵情况的发生，机场周围交通的不便利是造成大量旅客滞留在机场的主要原因之一。本系统可以通过对机场客流量时空分布的预测，从而实现机场及周边运力调配方案的优化。

本作品对将大数据预测应用在机场人流的场景中，连续动态且快速的预测了未来 10 分钟、20 分钟、30 分钟的人流分布，也可以预测未来两天的人流分布。基于此，机场可以与公交车、出租车等公共交通建立联系，完善运力调配平台，及时协调运力投入，同时发挥综合交通体系的功能，实现机场交通网状结构的搭建，确保旅客乘兴而来、满意而去。当系统预测到未来几天的客流量数据后，可以提前安排在相应客流量较大时间段内的运力调配方案，保证在相应时间段内旅客进出机场都可以搭乘到公交或出租车，防止旅客大量滞留机场情况的出现。在间隔较短的时间范围内，本作品可以实时监控各点的人流量，及时对运力调配方案做出相应调整，提高安排的准确性，机动调配运力，实现资源利用的最优化。

1.2 技术背景

1.2.1 数据处理

大数据时代到来，数据以爆炸式快速增长，尤其体现在互联网应用，电子商务等领域，其中最具有代表性的就是电子商务数据，淘宝网有 3.7 亿会员，在线商品 8.8 亿，每天生成约 20tb 数据[1]。此外，每个时间点都有人通过网络平台，发布个人文字信息、图片信息，其中 Facebook 每天至少生成 300TB 日志数据。Google 公司通过大规模集群和 MapReduce 软件，每月的数据量超过 400PB。而随着数据生成的自动化以及数据生成速度的加快，需要处理的数据量急剧膨胀，“大数据”已毋庸置疑成为当今研究热点。大数据的特点可归纳为 5 个“V”——Volume（大量）、Velocity（高速）、Variety（多样）、Value（低价值密度）、Veracity（真实性）。

因为大数据特点之一就是大数多样性，这就决定了经过各种渠道获取的数据种类和结构都非常复杂，给之后的数据分析处理带了极大的困难. 通过数据处理

与集成这一步骤，首先将这些结构复杂的数据转换为单一的或是便于处理的结构，为以后的数据分析打下良好的基础，因为这些数据里并不是所有的信息都是必需的，而是会掺杂很多噪音和干扰项，因此，还需对这些数据进行“去噪”和清洗，以保证数据的质量以及可靠性。常用的方法是在数据处理的过程中设计一些数据过滤器，通过聚类或关联分析的规则方法将无用或错误的离群数据挑出来过滤掉，防止其对最终数据结果产生不利影响；然后将这些整理好的据进行集成和存储，这是很重要的一步，若是单纯随意的放置，则会对以后的数据取用造成影响，很容易导致数据访问性的问题，现在一般的解决方法是针对特定种类的数据建立专门的数据库，将这些不同种类的数据信息分门别类的放置，可以有效地减少数据查询和访问的时间，提高数据提取速度。

那么，怎样构建数据仓库、怎样并行存储数据、怎样从如此庞大的数据中提取对后续有效发展的有利信息，怎样快速有效处理这些数据，已成为科研人员目前研究的重点。数据处理在这种背景下，应运而生。

数据处理是一种从大量但价值密度低的数据中抽取并推导出对于某些特定群体有价值，有意义数据的新兴技术，是系统工程和自动控制的基本环节。数据处理包括了对数据的采集、存储、检索、加工、变换和传输。

如今，数据处理在各领域都体现出了它的价值。在医疗行业，辅助医生通过对病人信息采集，对病情做出更科学。准确的判断，或对患者应季、频发病情提前预测并做出提醒；在能源行业，通过天气等信息处理，可预测最佳的太阳能或风力发电地址，提高能源效率；在通信行业，运营商通过对大数据的挖掘和处理，可以改善用户体验，未核实用户提高更优质的服务，优化网络质量，提高资源使用的效率；在零售业，通过处理商品与客户之间的活动数据，提供购买效率；在金融行业，可以帮助银行等金融企业分析客户的行为信息，掌握可发展的优质客户，实现客户利益最大化；在交通行业，对海量数据进行实时分析，可为出行人员提供精准、安全、通畅的可行性路线，扩大并行效果。

1.2.2 机器学习

机器学习是一门让计算机在非精确编程下进行活动的科学，也是一门多领域交叉学科，涉及概率论、统计学、逼近论、凸分析、算法复杂度理论等多门学科。

其核心点在于：计算机程序如何根据经验累积和数据分析来优化自身的性能标准。这恰巧揭示了数据背后的真实含义。大数据具有数据体量巨大、类型繁多、价值密度低等特性，如何从复杂、凌乱的大数据中完成数据的价值“提纯”成为大数据背景下，机器学习的一个亟待解决的问题。

机器学习[2]按照学习形式可分为监督学习和非监督学习；按照获取知识的表示形式可分为代数表达式参数、决策树、形式文法、产生式规则、框架和模式（schema）等十类不同形式；按照应用领域，又可分为专家系统、认知模拟、规划和问题求解、数据挖掘、网络信息服务、图象识别、故障诊断、自然语言理解、机器人和博弈等。虽然分类不尽相同，但机器学习领域的研究工作主要围绕三个方面进行：面向任务的研究、认知模型、理论分析。

机器学习虽然是人工智能研究中一个较为年轻的分支，但近年来，机器学习已经有了十分广泛的运用，例如：计算机视觉、数据挖掘、自然语言处理、搜索引擎、生物特征识别、医学诊断、检证券市场分析、测信用卡欺诈、DNA 序列测序、语音和手写识别、战略游戏和机器人运用。在运用到这些领域的同时，机器学习也从很多学科吸收了成果和概念，包括生物学、统计学、人工智能、哲学、信息论、认知科学、计算复杂性和控制等。

可以说，机器学习是人工智能应用的一个重要研究领域，也是人工智能和神经计算的核心研究课题之一。虽然现在的计算机系统和人工智能系统并没有很强的学习能力，但对机器学习的讨论和机器学习研究的进展，必将有效地解决现阶段不能应用机器学习来处理数据、满足科技和生产的问题，从而促使人工智能和整个科学技术的进一步发展 。

Spark 是一个高效的大数据处理平台用来提高数据挖过程的效率，掘随着 Spark 在大数据计算领域的暂露头角，越来越多的企业开始关注和使用。2014 年 11 月，Spark 在 Daytona Gray Sort 100TB Benchmark 竞赛中打破了由 Hadoop MapReduce 保持的排序记录。Spark 利用 1/10 的节点数，把 100TB 数据的排序时间从 72 分钟提高到了 23 分钟。

Spark 在架构上包括内核部分和 4 个官方子模块--Spark SQL、Spark Streaming、机器学习库 MLlib 和图计算库 GraphX。从 Spark 在伯克利的数据分析软件栈 BDAS（Berkeley Data Analytics Stack）中的位置。可见 Spark 专注

于数据的计算，而数据的存储在生产环境中往往还是由 Hadoop 分布式文件系统 HDFS 承担。

Spark 被设计成支持多场景的通用大数据计算平台，它可以解决大数据计算中的批处理，交互查询及流式计算等核心问题。Spark 可以从多数据源读取数据，并且拥有不断发展的机器学习库和图计算库供开发者使用。数据和计算在 Spark 内核及 Spark 的子模块中是打通的，这就意味着 Spark 内核和子模块之间成为一个整体。Spark 的各个子模块以 Spark 内核为基础，进一步支持更多的计算场景，例如使用 Spark SQL 读入的数据可以作为机器学习库 MLlib 的输入。大数据的挖掘是当今的研究热点，也具有很大的商业价值。传统方式在大数据 Hadoop 平台上利用 Mahout 以 MapReduce 的编程方式做数据挖掘，但是有一定的局限，比如效率较低。Spark 框架称为快数据，是基于内存的编程模型，它可以把中间的迭代过程不放在磁盘中，直接数据不落地在内存中执行，极大地提高了它的执行速度。因此，Spark 是大数据挖掘的新型利器。

从 Spark 的版本演化速度看，说明这个平台旺盛的生命力以及社区的活跃度。尤其在 2013 年来，Spark 进入了一个高速发展期，代码库提交与社区活跃度都有显著增长。以活跃度论，Spark 在所有 Aparch 基金会开源项目中位列前三。相较于其他大数据平台或框架而言，Spark 的代码库最为活跃，表现出强劲的发展势头。

下图为截止 2014 年 Spark 代码贡献者的增长曲线：

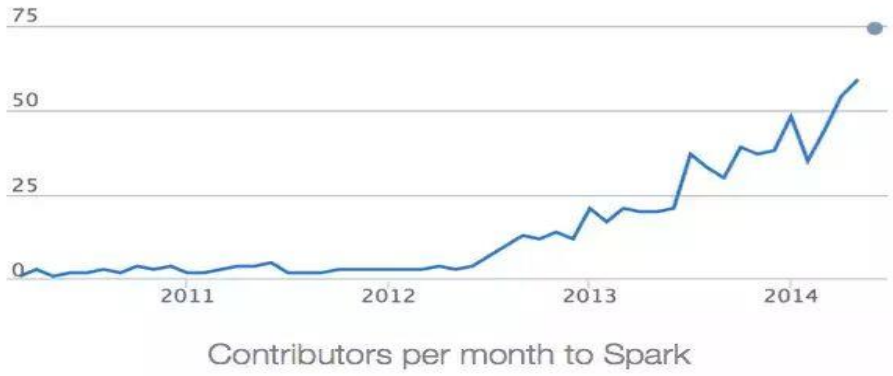


图 1.1 Spark 平台用户贡献图

从 2013 年 6 月到 2014 年 6 月，参与贡献的开发人员从原来的 68 位增长到 255 位，截止到 2015 年 6 月参与开发的人员已经达到 730 位（数据引用自 Spark Summit 2015 中报告），参与贡献的公司逐渐有来自中国的阿里巴巴、百度、网

易、腾讯和搜狐等公司。代码库的代码行也从 2014 年的 17 万行增长到 2015 年的 40 万行。

1.2.3可视化技术

种类繁多的信息源产生的大量数据，远远超过了人脑分析解释这些数据的能力。由于缺乏大量数据的有效分析手段，导致很多计算被浪费，这严重阻碍了科学研究的进展。为此，人们提出了一种有效的解决方法-----可视化。可视化技术作为解释大量数据最有效的手段而率先被科学与工程计算领域采用，并发展为当前热门的研究领域-----科学可视化。科学计算可视化[3]能够把科学数据，包括测量获得的数值、图像或是计算中涉及、产生的数字信息变为直观的、以图形图像信息表示的、随时间和空间变化的物理现象或物理量呈现在研究者面前，使他们能够观察、模拟、计算。

可视化技术是利用计算机图形学和图像处理技术，将数据转化成图形或图像在屏幕上显示出来，并进行交互处理的理论、方法和技术。它涉及到计算机图形学、图像处理、计算机视觉、计算机辅助设计等多个领域，成为研究数据表示、数据处理、决策分析等一系列问题的综合技术。目前，正在飞速发展的虚拟现实技术也是以图形图像的可视化技术为依托的。在互联网时代，可视化与网络技术结合使远程可视化服务成为现实，可视区域网络因此应运而生。

实时高效的数据处理固然重要，但是在较短的时间内对大量的数据和信息进行分析，是很难分析出数据变化的规律和得出一些结论。于是，这时候我们就要把大量的抽象数据转化为我们能够一目了然的线条变化趋势和图像，通过这样来清晰地得出结论。

可视化技术使人能够在三维图形世界中直接对具有形体的信息进行操作，和计算机直接交流。这种技术已经把人和机器的力量以一种直觉而自然的方式加以统一，这种革命性的变化无疑将极大地提高人们的工作效率。图形图像承载的信息量远多于语言文字，人类从外界获得的信息约有 80%以上来自于视觉系统。可视化借助于人眼快速的视觉感知和人脑的智能认知能力，可起到清晰有效地传达、沟通并辅助数据分析的作用。现代的数据可视化技术综合运用计算机图形学、图像处理、人机交互等技术，将采集或模拟的数据转换为可识别的图形符号。随着

狂交网络和移动互联网的兴起与发展，互联网、经济金融、社会公共服务等领域产生了一些特征鲜明的数据类型，主要包括文本信息、网络或图、时空数据及多维数据等。送些与大数据密切相关的信息类型与信息可视化的分类交叉離合，成为大数据可视化的主要研究领域。

可视化技术赋予人们一种仿真的、三维的并且具有实时交互的能力，这样人们可以在三维图形世界中用以前不可想象的的手段来获取信息或发挥自己创造性的思维。

1.3 政策背景

随着中国政府有关政策对促进通用航空发展的引导、保障和推进，中国通用航空市场规模不断扩大，市场活力正在被激发出来。

《通用航空飞行管制条例》修订版、《低空空域划设方案》、《通用航空飞行管制条例》修订版、《通用航空机场申报与审批管理程序》等都有力地鼓励了机场管理运营进行升级调整。

我国早已制定了多项政策推动通用航空发展。2010 年 11 月，国务院、中央军委印发《关于深化我国低空空域管理改革的意见》。《意见》提出通过 5 至 10 年的全面建设和深化改革，逐步形成一整套既有中国特色又符合低空空域管理规律的组织模式、制度安排和运作方式。具体实施分 3 个阶段，2011 年前为试点阶段；2011 年至 2015 年底前为推广阶段；2016 年至 2020 年为深化阶段。公开资料显示，2013 年 11 月，中国人民解放军总参谋部、中国民用航空局出台了《通用航空飞行任务审批与管理规定》。

国家又多项财政政策支持，可以加快培育通用航空市场，促进通用航空与运输航空协调发展；长期来看，通用航空的发展主要依靠市场机制解决，在条件成熟时，可以考虑从战略性新兴产业资金中安排一定资金，建立研发投入机制和用户补贴机制，加快飞机发动机、航电设备等核心技术的国产化，鼓励通航企业使用国产飞机，使通用航空产业成为继汽车产业之后又一个能大力拉动国内需求的新兴支柱产业。

第2章 产品、服务与技术

2.1 服务描述

由于机场的特殊性质，机场为旅客提供具有以下特性的服务：

安全性 对旅客精神、健康、生命安全以及行李的完整性和安全性的保障，对突发事件有能力解决。

功能性 机场能满足服务对象的基本要求，在正确的时间、地点，提供完整准确的相关信息和便捷、安全的服务。

效率性 在保证服务质量的同时，保证服务的效率。具体表现为，是旅客在最短的时间内，完成完整的基本流程。

舒适性 保持良好的环境卫生，把握相应的布局；机场人员有良好的精神面貌，服务过程中使旅客有亲切感。

2.2 产品技术

2.2.1 关键技术简介

本次作品的关键技术为数据挖掘技术[4]。其本意即为在大量的数据中挖掘有价值的信息，并以其作为基础，进行其他的科学研究。

由于人们的信息量增多，自动的数据收集工具和成熟的数据库技术导致大量数据存放在数据库、数据仓库，和其他信息存储中。我们握有大量有价值的数据，且不知如何应用，数据挖掘基础即在这种状况下应运而生。作为一种先进的技术，它随着信息量的增大，而逐步发展。

数据挖掘作为一个应用驱动的领域，吸纳了许多应用领域的技术，包括：统计学、机器学习、模式识别、数据库和数据仓库、可视化、算法、高性能设计等。在不同的领域，数据挖掘技术都给我们带来了同等的便利。在商业领域，我们根据统计大量的客户信息，找出具有相同特征的的客户群，确定客户的购买模式，以此来向不同客户群推送相关产品；在金融领域，我们借助数据挖掘，对现金流做分析和预测，提前为客户提供相应的资产管理方式，指导企业有规模的发展潜

在客户。在安全工程上，数据挖掘可以根据通话距离、通话时间或每天、每周的通话次数，分析偏离期望的模式，来提醒使用者是否收到电话欺骗等安全威胁。

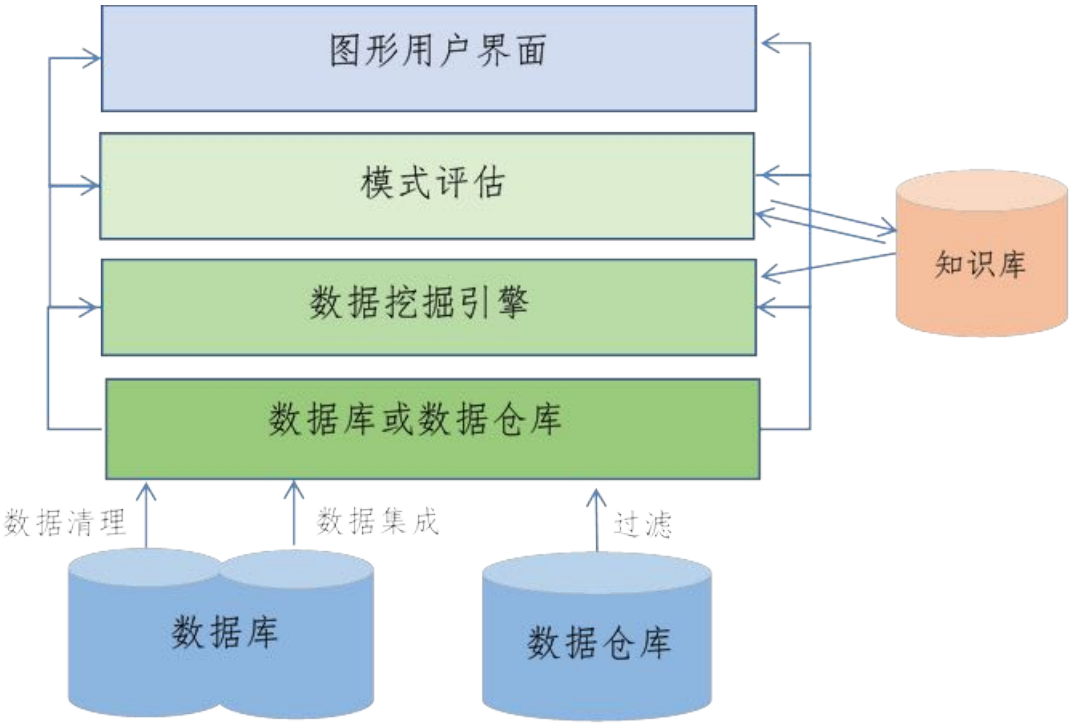


图 2.1 一般的大数据系统构建图

2.2.2 产品技术构架

本作品从数据采集开始，对数据依次进行分类、清洗等处理，供上层算法调用，最后在网页端提供可视化的界面系统作展示展示。上述架构可具体分为四层，如图 2.2 所示。

第一层为数据层，包含支持整个系统的数据，如 Wi-Fi 点连接人数、安检口旅客信息、航班信息等；

第二层为预处理层，将采集到的数据进行清洗，如删去异常值、添加缺失值等等，以供下一层使用；

第三层为分析处理层，具体是从清洗过的数据中提取有用的特征，按照区域分别建立子模型，最后将子模型加权融合实现我们预测系统的核心部分；

第四层为可视化界面层，提供一个网页端的可视化操作系统。

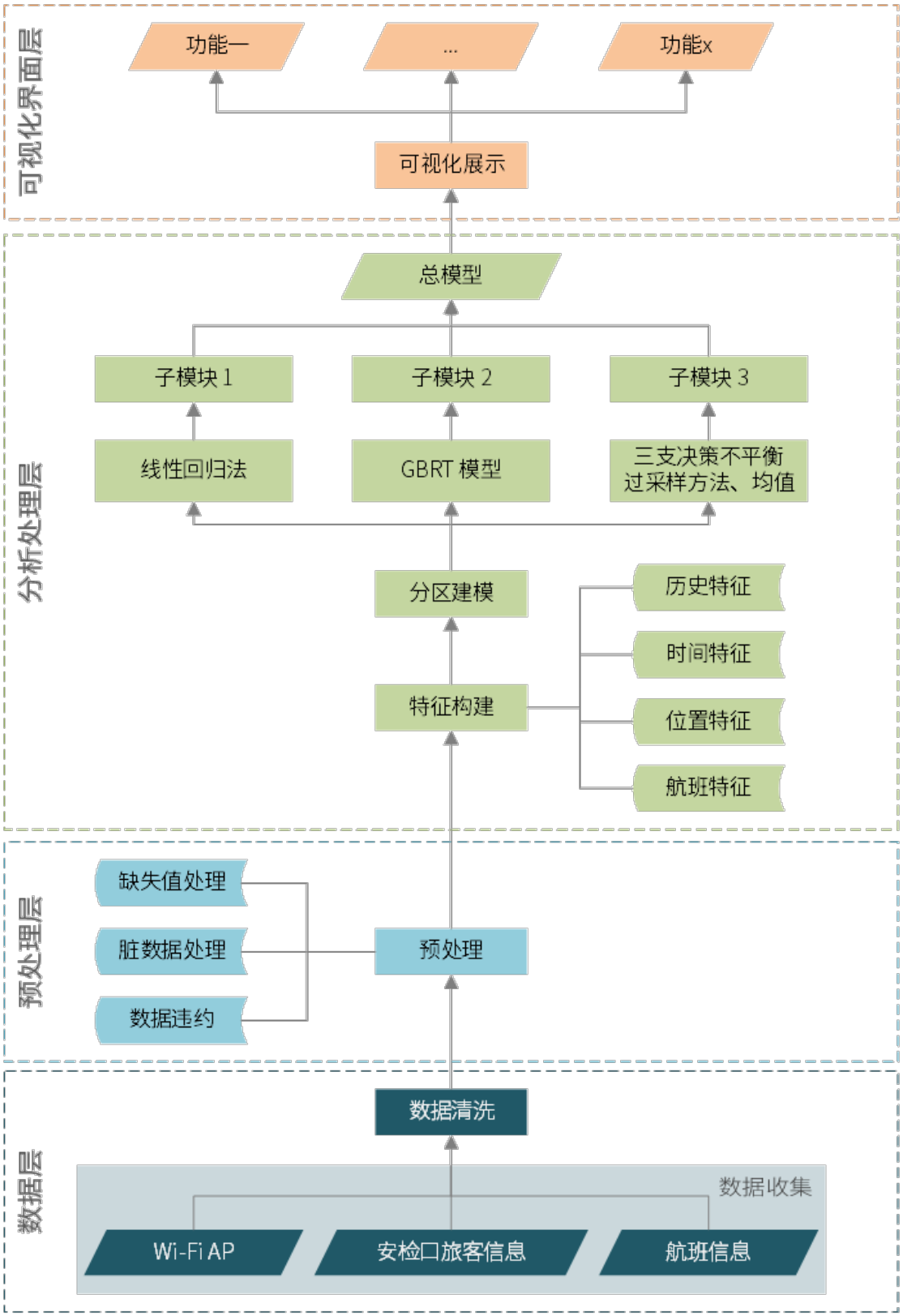


图 2.2 本系统架构图

可简化如下图：

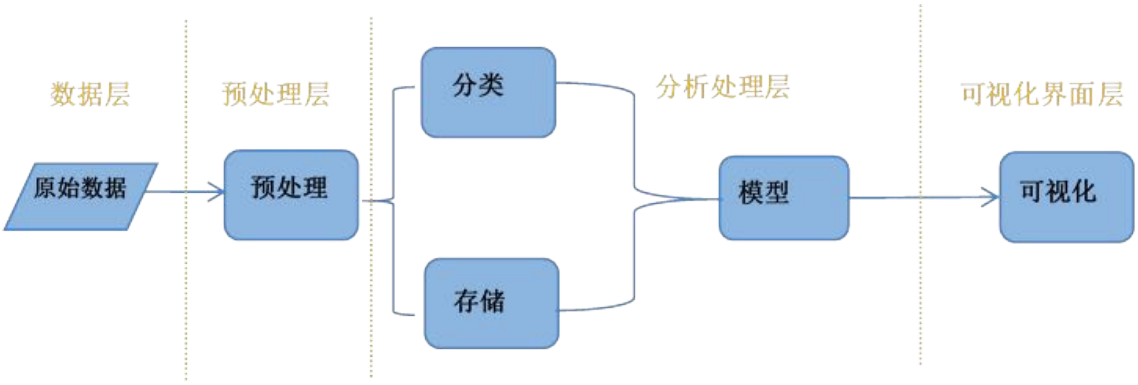


图 2.3 本系统架构简化图

2.2.2.1 数据层

数据层作为支撑整个系统的基础，此层存放机场所有 Wi-Fi 点的连接数据，登机口旅客数据，安检口旅客数据以及航班表的数据等，包括实时的旅客地点信息，机场提供的 Wi-Fi 点对应的登机口，安检口人数，飞机到达时间，飞机型号及载客量等等，这些数据均有可能作为算法的调用数据。

2.2.2.2 预处理层

数据预处理包括对从数据层获得的数据进行数据清洗，消除异常数据，填补遗漏缺失数据，降低噪声数据。同时，考虑到数据集大小对计算性能的影响，进行数据规约，使数据标准化，各指标得以处于同一数量级，适合被上层算法调用。除此之外，建立相应数据库，对处理好数据进行分类、存储。

2.2.2.3 数据分析层

数据分析层具体是从上一层已经清洗过的数据提取特征，特别要说的是这些数据中一些特征值相互关联不大，放在一起会起冲突反而影响预测结果，故我们按人流分布区域将数据分成三类，并针对这三类数据使用了回归模型在内的多种模型融合而成的算法来分别建模，最后将这三个子模型加权融合，实现我们预测系统的核心部分，对未来人流分布情况作出预测。同时，模型具有可优化型，可以再次更改，接受更多的特征，也可实现更多的预测功能。

2.2.2.4 可视化层

将分析预测得出的结论，通过可视化软件及相关代码，与后台数据连接，将其展示出来。并设以相应的查询、分析快捷方式，让使用者可以以简单、直观的方式得到后台产生的预测信息，了解当前机场人流分布情况、10 分钟乃至 2 天的机场人流预测分布情况、机场人流量是否过多应采取什么样的措施以及采取措施后有什么样的效果对比等等。

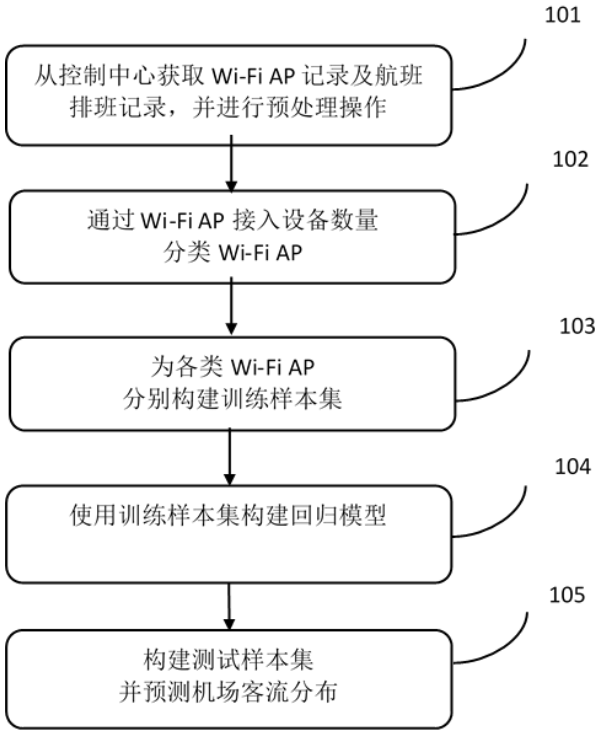
2.2.3 产品技术模型构建

本模型算法利用机场客流分布的相关特性，使用数据挖掘及机器学习的相关理论及方法，对机场的客流分布进行预测，达到有效利用机场资源，提升机场的生产运营效率。

我们利用了基于 Spark 内存计算大数据平台的 CLR 多标签数据分类方法。

可以把预测的每一天当成一个标签进行预测，10 天就有 10 个标签，或者把每一天的每一个十分钟当成一个标签进行预测。

基于机场 Wi-Fi AP 记录及航班排班记录的机场客流分布预测方法如下图，包



括：

图 2.4 模型算法步骤图 1

101 从控制中心获取 Wi-Fi AP 记录及航班排班记录，并进行预处理操作；

102 通过 Wi-Fi AP 接入设备数量分类 Wi-Fi AP；

103 为各类 Wi-Fi AP 分别构建训练样本集；

104 使用训练样本集构建回归模型；

105 构建测试样本集并预测机场客流分布。

作为本算法的进一步改进，具体步骤为：

步骤 1：从控制中心获取 Wi-Fi AP 记录及航班排班记录，并选取最近 30 天的记录。

步骤 2：对步骤 1 获取的 Wi-Fi AP 记录及航班排班记录进行缺失值处理。对于某一 Wi-Fi AP 的缺失数据，使用与缺失数据最相近的 15 天记录对应时刻设备连接数量的均值进行填充。

步骤 3：对步骤 2 处理后的 Wi-Fi AP 记录进行脏数据处理。使用 ARMA 模型对数据进行平滑处理。

步骤 4：对步骤 3 处理后的 Wi-Fi AP 数据进行数据规约。以 10 分钟为单位对 Wi-Fi AP 连接数以平均值进行规约，即每 10 分钟生成一条数据。计算方法如公式⑴所示，其中，x_{ij} 为某 Wi-Fi AP 第 i 个十分钟的第 j 个分钟的设备连接数量，r_i 为该 Wi-Fi AP 规约后的第 i 个十分钟 的设备连接数量。

$$r_i = \frac{\sum_{j=0}^{9} x_{ij}}{10}$$

公式⑴

步骤 5：对于各个 Wi-Fi AP，计算其设备连接数的方差，并根据其方差由大到小进行排序，然后使用二八法则划分 Wi-Fi AP 为两类。方差较小的 Wi-Fi AP 为第一类 Wi-Fi AP，方差较大的 Wi-Fi AP 为第二类 Wi-Fi AP。

步骤 6：对于第一类 Wi-Fi AP，取最近 15 天的数据，建立第一类 Wi-Fi AP 训练集。

步骤 7：对于第二类 Wi-Fi AP，使用预测日前 15 天的数据进行标签提取，标签为某一时刻该 Wi-Fi AP 的设备连接数。

步骤 8：对第二类 Wi-Fi AP 进行特征提取。选择标签日前 15 天的记录进行特征提取，其特征包含 3 部分：

（1） 历史特征：对于该 Wi-Fi AP 的同一时刻，分别计算该 Wi-Fi AP 在以天为单位的同一时刻的均值、最小值、最大值和方差信息。

（2） 航班特征：航班是影响连接数波动的主要因素之一，根据航班排班记录的登机口位置信息，统计该登机口位置 10 分钟、30 分钟、60 分钟及 120 分钟内飞机起飞数量，并与 Wi-Fi AP 的位置信息关联后进行数据合并。

（3） 位置特征：包含 Wi-Fi AP 所在的区域、所在楼层、所在组编号和 Wi-Fi AP 坐标信息。

步骤 9：对于第一类 Wi-Fi AP，取步骤 6 所建立的第一类 Wi-Fi AP 训练集，按公式⑵构建第一类 Wi-Fi AP 回归模型。其中，y_{ij} 为编号 i 的 Wi-Fi AP 的 j 时刻的预测值，set1 为第一类 Wi-Fi AP 集合。y_{ij} 由公式⑶得出，其中，x_{ijk} 为编号 i 的 Wi-Fi AP 第 k 天 j 时刻的设备连接数量[14]。

$$Y_1 = \bigcup_{i \in set1} y_{ij}$$

公式⑵

$$y_{ij} = \frac{\sum_{k=1}^{15} x_{ijk}}{15}$$

公式⑶

步骤 10：对于第二类 Wi-Fi AP，其特点为设备连接数的方差较高。对于这类 Wi-Fi AP，使用公式⑷构建第二类回归模型。其中，y_{ij} 为编号 i 的 Wi-Fi AP 的 j 时刻的预测值，set2 为第二类 Wi-Fi AP 集合。y_{ij} 由公式⑸得出，其中，x_{ij} 为测试样本，h 函数为使用 S1032 所构建的第二类训练集所训练的基于最优叶子分裂的 GBDT 回归模型。

$$Y_2 = \bigcup_{i \in set2} y_{ij}$$

公式⑷

$$y_{ij} = h(x_{ij})$$

公式⑸

步骤 11：对第一类模型和第二类模型进行集成，集成方法如公式⑹所示。

$$Y = Y_1 \bigcup Y_2$$

公式(6)

步骤 12：对于第一类 Wi-Fi AP 集，使用第一类模型进行预测，记其预测结果为 P1。

步骤 13：对于第二类 Wi-Fi AP 集，使用步骤 8 进行特征提取[15]，并使用第二类模型进行预测，记其预测结果为 P2。

步骤 14：对第一类模型预测结果和第二类模型的预测结果的集成作为最终预测结果，集成方法如公式(7)所示。

$$P = P_1 \bigcup P_2$$

公式(7)

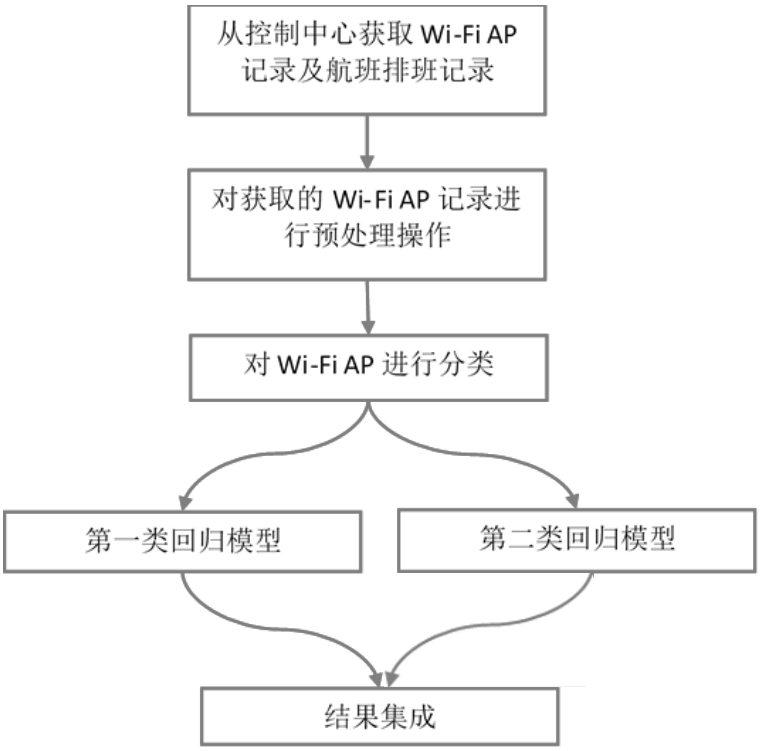


图 2.5 模型算法步骤图 2

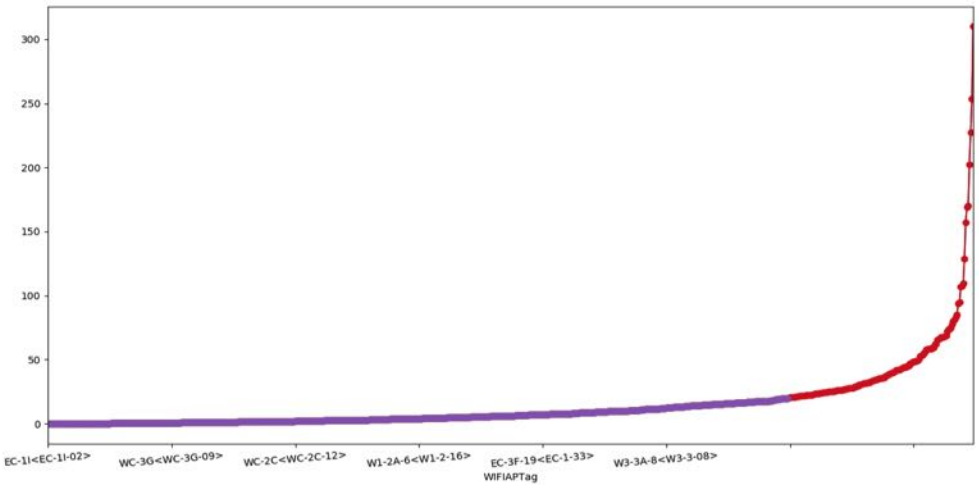


图 2.6 Wi-Fi 点特征数量

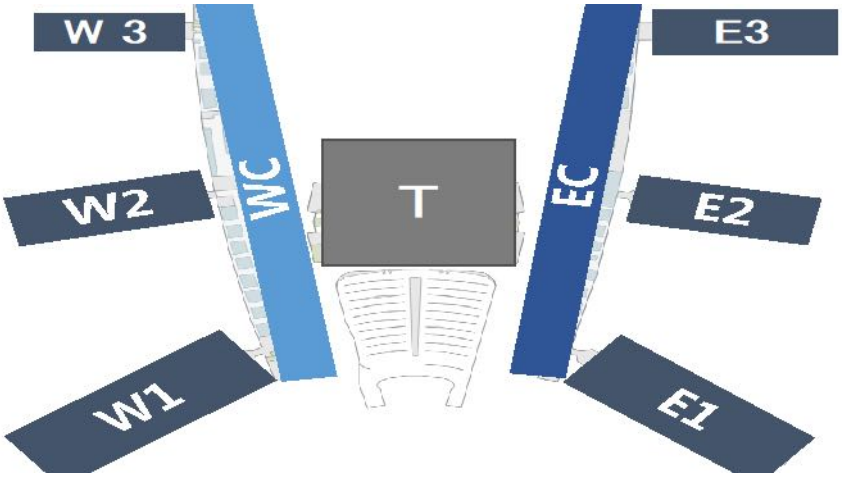


图 2.7 机场分区图

我们观察到可以将白云机场分为四个区域：登机区域 （W1、W2、W3、E1、E2、E3）、T 区域、WC 区域与 EC 区域。

对于候机区域的无线 AP 点，只需要考虑与之距离最近 6 个登机口的起飞航班特征；对于 T 区域，只需要考虑 W、E 这两个大区域的总起飞航班特征及 T 区域内 AP 组号；对于 WC 区域，只需要考虑 W1、W2、W3 这三个登机区域的总起飞航班特征及 WC 区域内 AP 组号；对于 EC 区域，只需要考虑 E1、E2、E3 这三个登机区域的总起飞航班特征及 EC 区域内 AP 组号。另外，将 B901-B909 的航班起飞特征放在 WC 区域，因为这 9 个登机口位于 WC 区域的一楼。将单个回归模型分为四个子模型后——从单个回归模型预测所有区域到为不同区域分别建模，输入的维度大幅度降低。

另外，对于时间类特征，如星期、天内分钟偏移等，我们将其看作 1 维整数，

而不使用 one-hot 高维编码，这一技巧进一步降低了输入维度。与线性回归等参数回归模型不同，GBRT 是基于空间划分的回归模型，对于诸如星期这类有序变量，它可以根据训练数据将其合理地划分为多个区间，比如可以将星期划分为[星期一, 星期四]、[星期五, 星期日]两个区间。

最后，对于 11 日凌晨的 3 个小时（00:00-02:59），通过为它们引入 10 日夜晚的近期信息，比如 10 日最后一个时间点的连接数及安检信息等，可以进一步优化预测结果。也就是说，预测短期内客流分布可以做得更准确。

在该模型中运用了机器学习中的很多算法。比如：线性回归算法，GBRT 算法，均值算法等多种算法，它们相互融合从而提高预测能力[16]。

2.2.3.1 线性回归算法模型

线性回归(Linear Regression)是利用称为线性回归方程的最小平方函数 对一个或多个自变量和因变量之间关系进行建模的一种回归分析[17]。这种函数是一个或多个称为回归系数的模型参数的线性组合。

机器学习的回归思想就是通过丢给机器学习和观察数据来找到杂乱数据间隐藏的规律，通过建模和算法使得规律得到的结果不仅与真实的结果越逼近越好，而且在新的数据上也有很好的预测准确性，也就是好的泛化能力[18]。

2.2.3.2 GBDT 算法模型

迭代决策树算法，由多棵决策树组成，所有树的输出结果累加起来就是最终答案。它在被提出之初就和 SVM[19]一起被认为是泛化能力（generalization)较强的算法。近些年更因为被用于搜索排序的机器学习模型而引起大家关注。

GBDT 是回归树，不是分类树。它是用于预测实数值。其核心就在于，每一棵树是从之前所有树的残差[20]中来学习的。为了防止过拟合[21]，和 Adaboosting 一样，也加入了 boosting 这一项。

2.2.4 详细设计

2.2.4.1 数据预处理

从控制中心获取 Wi-Fi AP 记录及航班排班记录，并选取最近 30 天的记录。对获取的 Wi-Fi AP 记录进行预处理操作，所述的预处理操作具体为：

（1） 缺失值处理：对于某一 Wi-Fi AP 的缺失数据，使用与缺失数据最相近的 15 天记录对应时刻设备连接数量的均值进行填充。

（2） 脏数据处理：使用 ARMA 模型对数据进行平滑处理，对于单一 Wi-Fi AP 记录的设备连接数量，经过 ARMA 模型处理后的数据变化平稳，有效的降低了噪声数据。同时脏数据处理的时候可以进行离群点检测，快速发现离群点（即噪声数据）。

我们在此提出了一种基于内存计算 Spark 大数据平台的 OPTICS 算法**【已申请专利】**，具体包括：读取大规模数据集，创建分布式数据集 RDD，完成初始化操作；对数据结构进行并行划分，得到最优数据集分区；根据最优数据集分区生成的 RDD，并行计算邻居样本数量和核心距离，对每个分区并行执行 OPTICS（通过点排序识别聚类结构）算法可以得到每个分区的簇排序并持久化存储；通过簇排序给每个分区赋予簇号，合并分区，使每个样本得到全局的簇号。

本发明利用 Spark 分布式并行技术，找到数据结构的最优划分结构，并行计算得到每个分区的簇排序，通过 OPTICS 算法的簇排序[6]，用户可以从不同层次结构观察数据集的内在聚类结构。本发明具体包括以下步骤：从分布式文件系统 HDFS（Hadoop Distributed File System，Hadoop）上读入数据集，并创建一个 Spark 的上下文 SparkContext（程序运行初始环境）对象，利用对象的抽象数据结构函数 parallelize(DataSet)或 textFile(DataSet URL)创建分布式数据集 RDD，创建完成的分布式数据集可以被并行操作。 把 RDD 中每个样本通过 map()（每行数据执行同一操作的函数）函数转换为对应的自定义类 Point（用来存储每个样本的各种信息），Point 类中存储有每个样本的值及其相关信息。可以输入聚类初始的半径 ε 和半径内最小的邻居数 MinPts。

对 RDD 进行划分得到数据集的最优结构，具体可采用如下方法：

（1） 通过 RDD 的行动函数（reduce()或者 fold()（一种通过折叠 RDD 来计

算整个 RDD 信息的函数））计算数据集的所有维度,得到维度数 N（一般维度差异最大的有 5 个维度），广播整个数据集。

（2） 把 RDD 分成 N 个分区，每个分区能够获取到前面的广播变量，每个分区分别根据相关维度，各自生成树形结构。树的每个节点都是一个盒子（box 类），盒子有前边界和后边界及存储包括的样本数组。生成树形结构时，按照维度进行数据集的划分。首先按照维度将数据集等距离进行平分，形成两个盒子，这两个盒子再进行等距离平分，直到盒子的样本数据个数小于设定值或者盒子的前边界值减去后边界值小于 2× ε 。标记每个样本属于的盒子，或是否属于每个盒子的前后边界。标记每个样本与所属盒子的前后边界的关系。例如某个样本点到所属盒子的前边界线的距离小于 ε，则为此盒子的前边界点，若到所属盒子的后边界线的距离小于 ε，则为此盒子的后边界点。

（3） 遍历所有分区的划分树，得到盒子数组。调用 Spark 平台所提供的累加变量，汇集所有分区盒子数组的情况，寻找到盒子数组划分均匀并且盒子数组长度最多的分区，得到最优分割分区。具体可采用如下方法：每个分区都获取自己分区的盒子数组的总长度 L，每个分区的对应的盒子数组中样本个数最多的样本数 M，盒子数组的平均样本数 N，根据公式 P=N/M 计算判断倾斜的标准[7]。每个分区的 L 和 P 标准化，取 L+P 值最大的分区为最优分割分区，保存最优分割分区，去掉其余分区。保存的最优分割分区其内部的盒子数组就是最优划分结构的盒子数组。

进一步地，可根据最优分割分区的最优划分结构的盒子数组得到新的 RDD，并计算样本邻居和样本点的核心距离，具体可包括如下方法：

（1） 首先广播整个盒子数组（即把盒子数组分发到每台机器并存储一个副本,可理解为每个分区都会得到一个盒子数组的副本，减少大幅度任务计算的时间），另外再单独把盒子数组生成为 RDD 结构，RDD 每个分区分别获得序号相对应的盒子。例如 RDD 的 0 号分区获得盒子数组的 0 号盒子，依次类推。因为盒子数组和 RDD 的盒子本身储存的样本点都是一样，广播处理的话可加快计算的速度。

（2） 每个 RDD 分区中的盒子分别与广播的盒子数组中其序号对应的盒子

及其序号前后盒子进行计算样本邻居。例如 0 号分区与盒子数组的 0 号和 1 号盒子进行计算，1 号分区与 1 号，0 号和 2 号盒子进行计算，依次类推。因为每个盒子存储的都是大量的样本点，分区中的盒子与其序号对应的盒子进行计算样本点与样本点之间的距离，可得到一个样本点周围有哪些样本点是其邻居样本点。而分区中盒子与其序号对应前后的盒子计算点与点之间欧几里得距离是为了计算分区盒子中边界点和前后盒子中边界点是否是邻居关系。是否是邻居样本点的判断标准取决于用户设定的阈值。

（3） 根据邻居样本点及与其目标样本点的欧几里得距离[8]，根据如下公式(1)计算得到每个样本的核心距离。

$$core-dist_{\varepsilon,MinPts}(p)=\begin{cases}UNDEFINED & if|N_{\varepsilon}(p)|\prec MinPts\\ MinPts-th\ \ smallest\ \ distance\ \ to\ \ N_{\varepsilon}(p) & otherwise\end{cases}$$

(1)

$core-dist_{\varepsilon,MinPts}(p)$ 表示对于样本点 P 的核心距离，其中 ε 表示样本点 P 的邻域半径，MinPts 表示使得样本点 P 成为核心样本点的最小邻居样本点个数，core-dist 表示核心距离。

$if\ |N_{\varepsilon}(p)|\prec MinPts$ 表示当样本点 P 周围的邻居样本点个数小于设定阈值个数 MinPts 时，其中 $|N_{\varepsilon}(p)|$ 表示样本点 P 周围的的邻居样本点个数。

$MinPts-th\ \ smallest\ \ distance\ \ to\ \ N_{\varepsilon}(p)$ $UNDEFINED$ 表示此时 P 的核心距离无定义。

表示使得样本点 P 周围至少有 MinPts 个邻居样本点的最小邻域半径，其中 MinPts-th smallest distance 表示满足 MinPts 邻居样本点个数最小半径距离。

根据公式（2）计算每个样本的可达距离：

(2)

$$reachability-dist_{\varepsilon,MinPts}(o,p)=\begin{cases}UNDEFINED & if|N_{\varepsilon}(p)|\prec MinPts\\ Max(core-dist_{\varepsilon,MinPts}(p),dist(p,o)) & otherwise\end{cases}$$

$reachability-dist_{\varepsilon,MinPts}(o,p)$

表示样本点 O 到样本点 P 的可达距离，reachability-dist 表示可达距离

$Max(core-dist_{\varepsilon,MinPts}(p),dist(p,o))$ 表示取样本点 P 的核心距离和样本点 P

和样本 O 的欧几里得距离中最大的一个，其中 dist 表示 distance，欧几里得距离。

按照最优结构划分并计算距离得到每个分区中每个样本点（Point 类）的邻居样本点，并把邻居样本点的序号存储到样本点（Point 类）后，每个分区进行经典 OPTICS（通过点排序识别聚类结构）算法[9]，对所有样本的可达距离进行排序，每个分区得到簇排序，调用 saveAsTextFile（“outpath”）（把 RDD 存储到 Hadoop 分布式文件系统的函数）函数将簇排序持久化存储。每个分区的簇排序只能进行数据结构层面上观测作用，而为了得到每个样本点的最终簇号，需要输出聚类结果。所以根据每个分区的簇排序，可以依照用户的需求在不同层次结构上得到簇号，然后合并分区，输出聚类的结果。具体包括：

（1） 根据用户输入的判断距离阈值 B，从每个分区的簇排序中按顺序提取样本，首先先设定一个初始类别，如果该样本的可达距离不大于 B 则属于当前正在判定的类别。如果样本的可达距离大于 B 且核心距离大于 B，则为噪声[10]。如果样本的可达距离大于 B 且核心距离小于 B，则属于下一个新的类别，则另起一个新的类别。

（2） 生成全局合并簇号的 Map（键值对应的数据存储结构）[11]，保留每个分区的边界样本形成盒子边界数组，广播该数组。按每个分区的前边界点和数组序号对应的前一个盒子的后边界点进行循环，如果存在一对样本的距离小于给定 B，则加入 Map 中。

（3） 每个分区根据该 Map 合并簇号，并输出最终的聚类结果，如不符合期望公式，期望公式:已分配簇号的样本数/总样本数>=0.8(阈值 0.8 可以根据用户的实际情况进行调整)[12]，则可以返回步骤（1），直到符合期望公式为止。

$$r_i = \frac{\sum_{j=0}^{9} x_{ij}}{10}$$

数据规约：考虑到数据集大小对计算性能的影响，按公式⑴对数据进行规约。

其中，x_{ij} 为某 Wi-Fi AP 第 i 个十分钟的第 j 个分钟的设备连接数量，r_i 为

该 Wi-Fi AP 规约后的第 i 个十分钟的设备连接数量。

同时，传统机器学习通常以总体最大分类精度为目标，这一目标必然会导致算法提高多数类样本的分类精度，而忽略样本集中小样本的预测精度，使得分类器性能大幅度下降，得到的分类器具有很大的偏向性．本属于稀有类的样本往往被错分到大类．使得少数类样本的分类精度达不到人们的需求。因此，如何有效地提高少数类的分类准确率和分类器的整体性能已成为数据挖掘领域的一个热点。我们用了数据预处理的解决方法，目的是降低类别之间的不平衡性，在此层面上主要的方法是重采样，增加小类样本的数目（过采样）或减少大类样本的数目（欠采样），**此算法也已申请专利（见附件专利 2）。**

2.2.4.2 特征构建

2.2.4.2.1 历史特征

除了显示连接 Wi-Fi 的人数统计，无线 AP 的同一时间点历史连接数据也可以大致推算出当前时刻该无线 AP 的连接数。从图 2.8 可以看出，真实值一直围绕着历史均值上下浮动，而三种历史均值在不同情况下各有优势。我们以天为单位，计算了包括 AVG、MIN、MAX、STDDEV 多个指标，计算方式有同点、同时段、同区域三种，时间窗口为最近 1/2/3/4/5/6/7/14/30 天。在构建中，我们考虑加入了同时段、同无线 AP 的历史数据与同时间点、同区域内无线 AP 的历史连接数据，以降低噪音数据。

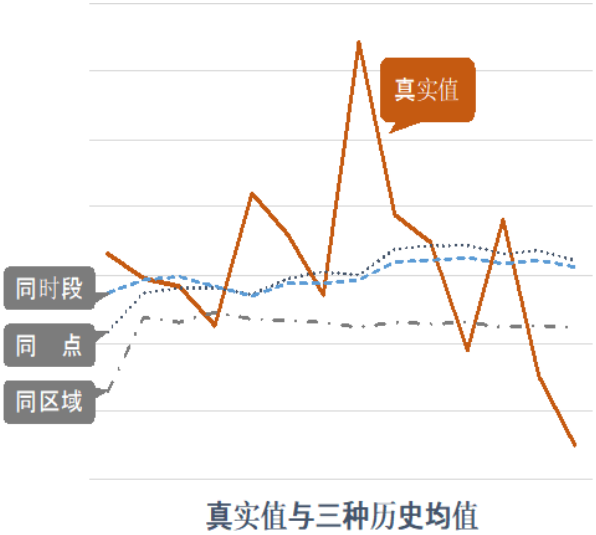


图 2.8 机场 Wi-Fi 连接对比图

2.2.4.2.2 航班特征

时序的历史特征可以为我们提供较为准确的人流信息，但是在机场由于天气原因，总会有航班晚点或延误的意外发生，航班是影响连接数波动的主要因素之一，所以需要刻画航班特征来是数据更为精准。根据航班排班记录的登机口位置信息，我们统计该登机口位置 10 分钟、30 分钟、60 分钟及 120 分钟内飞机起飞数量，并与 Wi-Fi AP 的位置信息关联后进行数据合并。

由于旅客的乘机时间不同，到达机场的时间也不同，所以我们将其分时间段进行统计。客机是否能准时起飞，对在机场等待的游客，包括接机区、登机区甚至是等待取行李的人流量有很大的影响。如果客机不能准时起飞，会增大在机场的人流密度，占用有限机场的设施，是一种资源浪费。

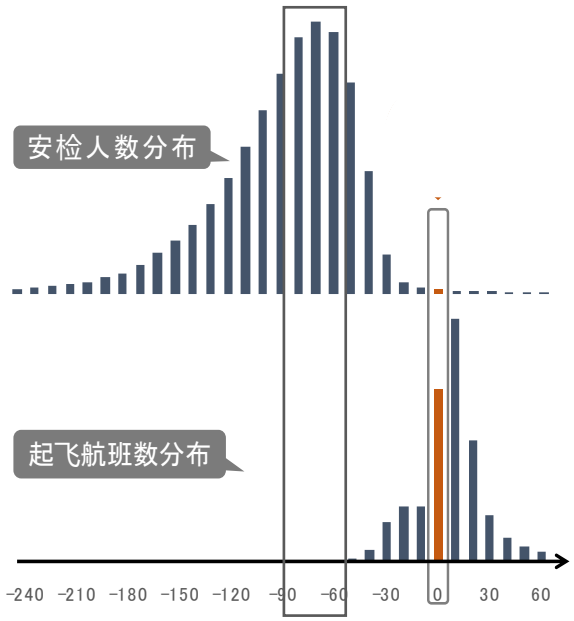


图 2.9 安检人数与航班对比图

2.2.4.2.3 位置特征

位置信息包含 Wi-Fi AP 所在区域、所在楼层、所在组编号及 Wi-Fi-AP 坐标信息，可以帮我们辨别特殊位置，类似于安检口、商店、登机口等。由于在机场中分布的 Wi-Fi 接入点地点不同，某一点的 Wi-Fi 接入量增加，即代表这一点处的人流增加，同样掌管着一点的设备正在被更多的人使用。所以对于特殊的位置，人流流量会有明显的不同。对于特殊位置和一般位置，需要标记特征，在连入数

据库的时候要使用均值对数据进行处理。

除此之外，机场的特殊位置，也包括机场大巴、连接机场的轻轨站，及周围的交通工具。有利的位置特征，能帮助我们准确判断出此处的人流量，是旅客们能够有序、不停留的乘坐设施，以免造成交通拥堵现象。

2.2.4.2.4 时间特征

时间信息用于帮助我们分辨特殊的时间点、时间段。在一天之中，以早 8：00 至下午 3：00 的飞机航班最多，而夜间的人流量就相对较少。除此之外，周末、节假日的人流量较平时有明显上升。所以我们截取每十分钟的人流量作为数据基础。将节假日出现的特殊人流量做数据规约，对数据做平滑处理[13]。并在遇到异常人流量的预兆时，将此信息提前反馈到可视化界面上，对机场做出提醒。

2.3 产品展示及使用

2.3.1 使用范围

“A-guardian”系统基于大数据分析，能提供预测、分析与合理的解决方案，可用于机场、火车站、交通要道等客流量大且流动频繁的场合，不仅面向场地管理者，而且面向旅客用户，此处展示以广州白云机场为例。

2.3.2 产品说明

本产品以广州白云机场为研究场景，通过分析白云机场 Wi-Fi 数据和安检登机值机数据来构建客流量预测模型，实现对机场航站楼客流量的准确预测，聚焦机场航站楼客流量预测和机场停机位资源分配优化这两大机场业务难题，用大数据技术支撑机场业务的快速发展。

2.3.3 运行环境

2.3.3.1 硬件

本系统界面放在网页端，所以通常的电脑，手机，只要是能连接互联网的平台均适用于本系统。

2.3.3.2 软件

常见 Web 浏览器（IE、FireFox、Chrome 等），桌面端（Windows、Linux、macOS 等）。

2.3.4 系统界面及功能介绍

2.3.4.1 登陆与注册

用户登录

用户连接机场 Wi-Fi 后选择 A-guardian 服务，或输入指定网址即可跳转到以下界面：

图 2.10 系统连接界面

如果用户已经注册则点击 Login 登录:

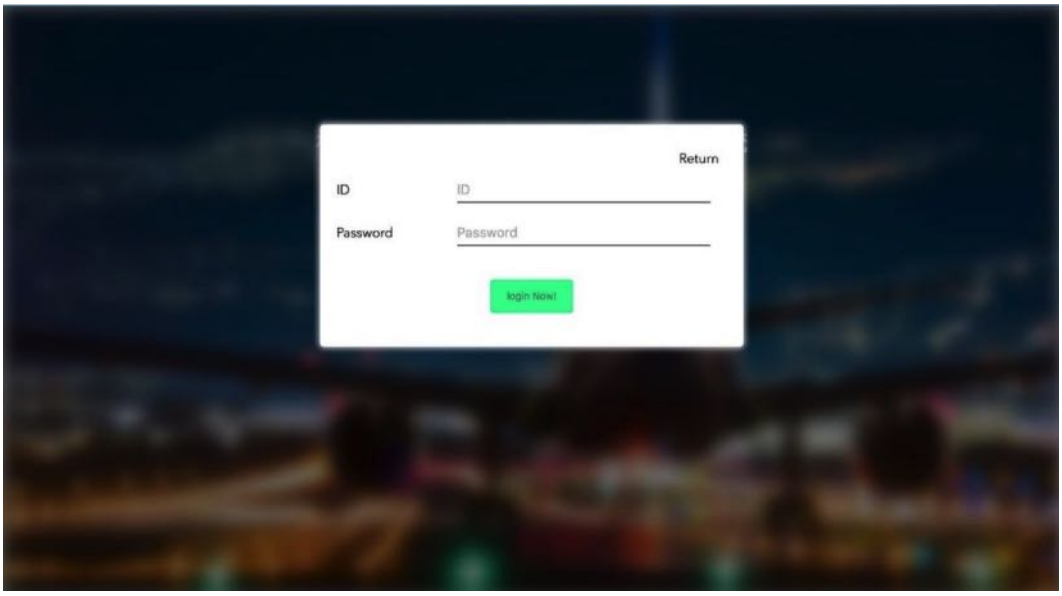


图 2.11 用户登陆界面

用户注册

如果用户未注册则点击 Register 注册

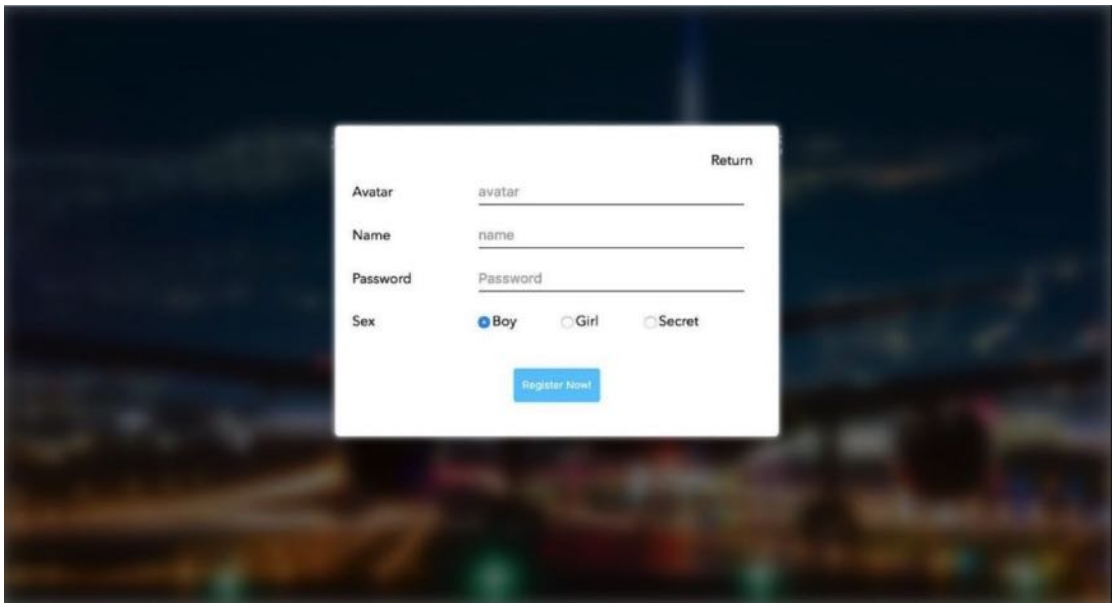


图 2.12 用户登陆界面

2.3.4.2 系统界面功能展示

2.3.4.2.1. 系统主界面

用户登陆后进入系统主界面，该界面显示机场实时的客流分布热度图，除此之外，界面下方还有一个时间条，可以拖动选择未来 10 分钟，20 分钟……1 天以及 1 天后的人数热度图预测。在界面最上方有控制面板、系统控制和数据统计三个选项。（注意，热度图显示需要点击上方菜单栏“系统控制”中的“启动系统”，否则只显示机场地图）

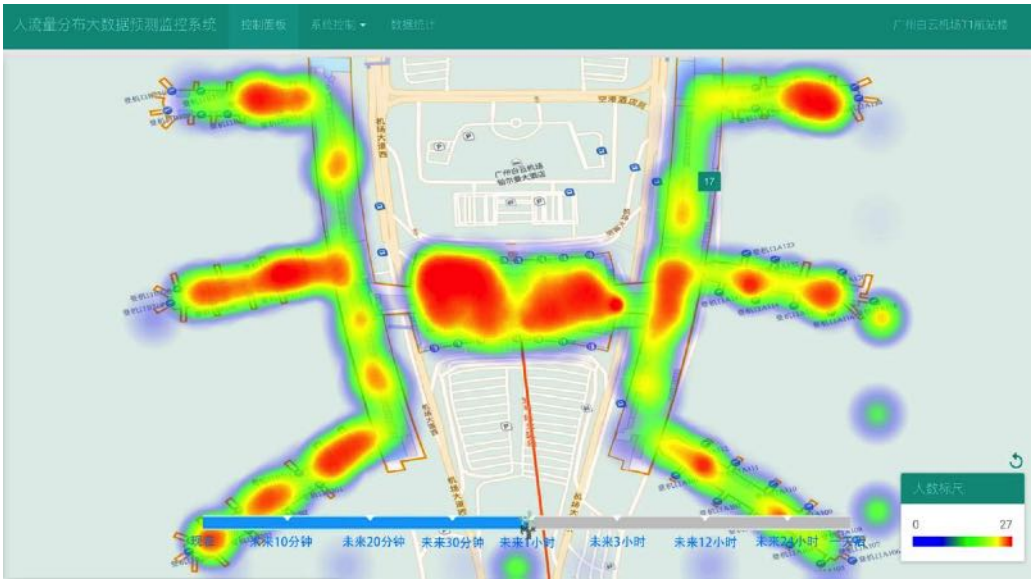


图 2.13 系统主界面

2.3.4.2.2. “控制面板”功能

2.3.4.2.2.1. “控制面板”主功能

用户点击控制面板后子界面弹出,子界面包括机场各区域的实施人数，如果超出平均值，系统会有 warning 的提示，提醒旅客到人数少的区域等待，提醒工作人员做出相应处理，当某区域人数大幅超出 warning 值并仍然急速增加时，系统会给出 danger 警告。

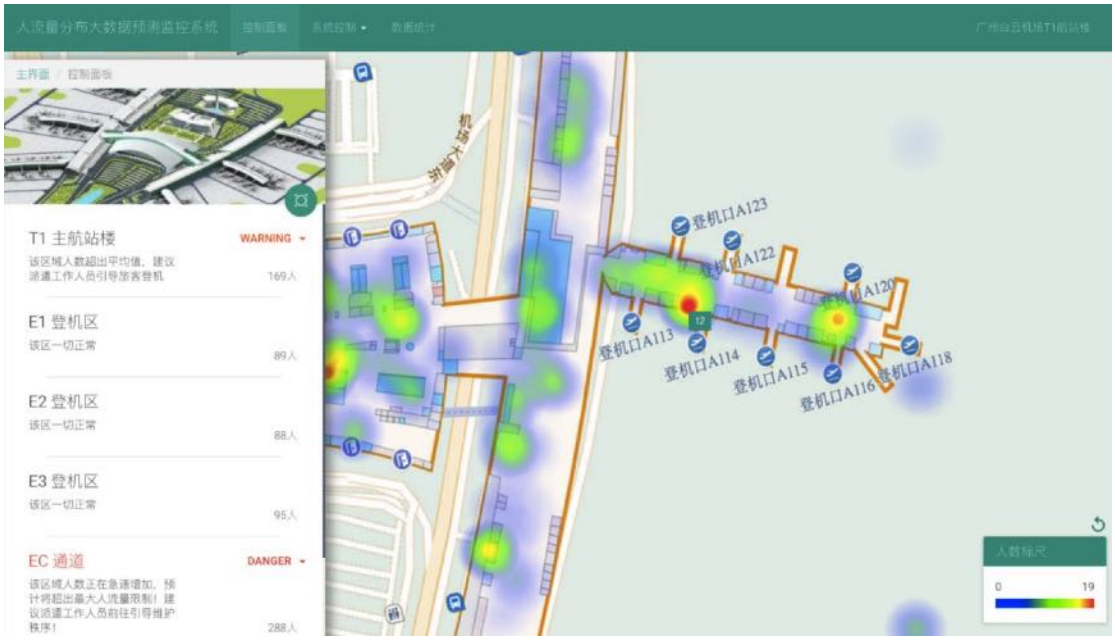


图 2.14 控制面板功能图

子界面中相应的区域对应如下：

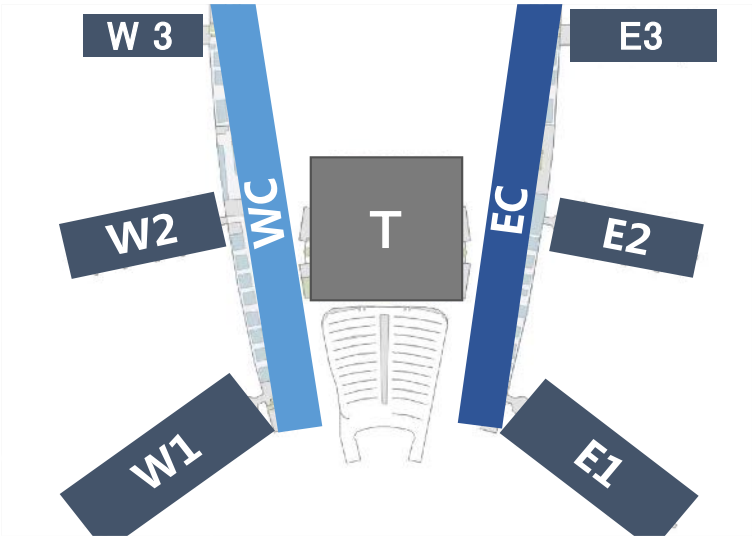


图 2.15 机场对应区域

2.3.4.2.2.2. “控制面板”细化功能——提供优化建议

当机场工作人员点击 warning 或 danger 标签时会弹出相应的建议措施，比如：向警卫部发送该警告、通知塔台、提升预警等级和忽略此条预警。

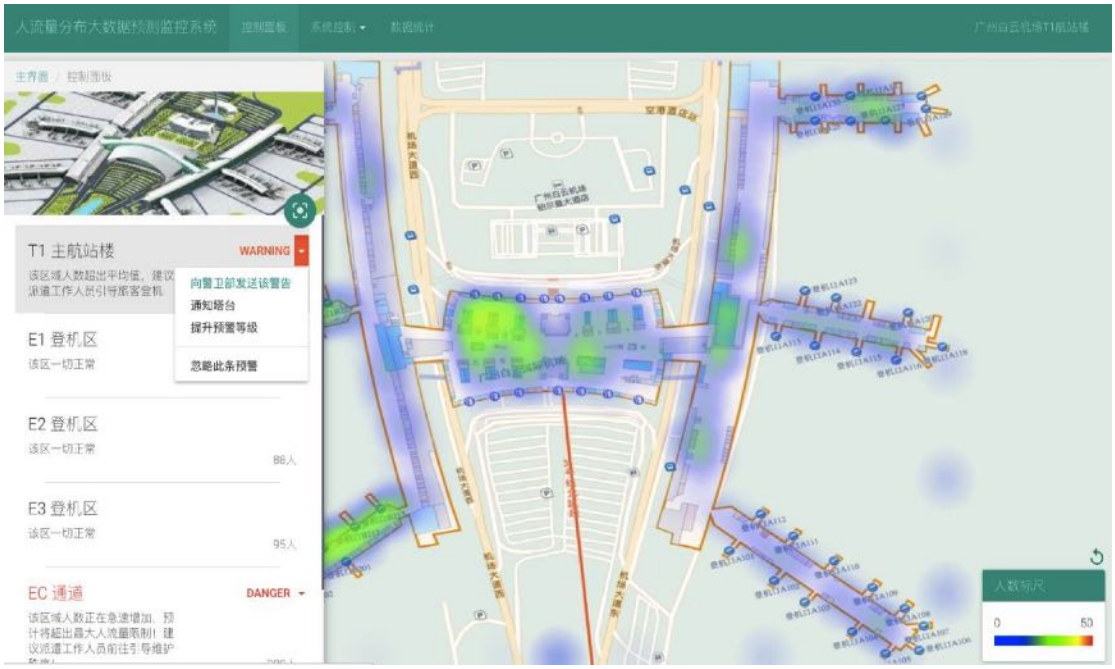


图 2.16 warning 警告措施处

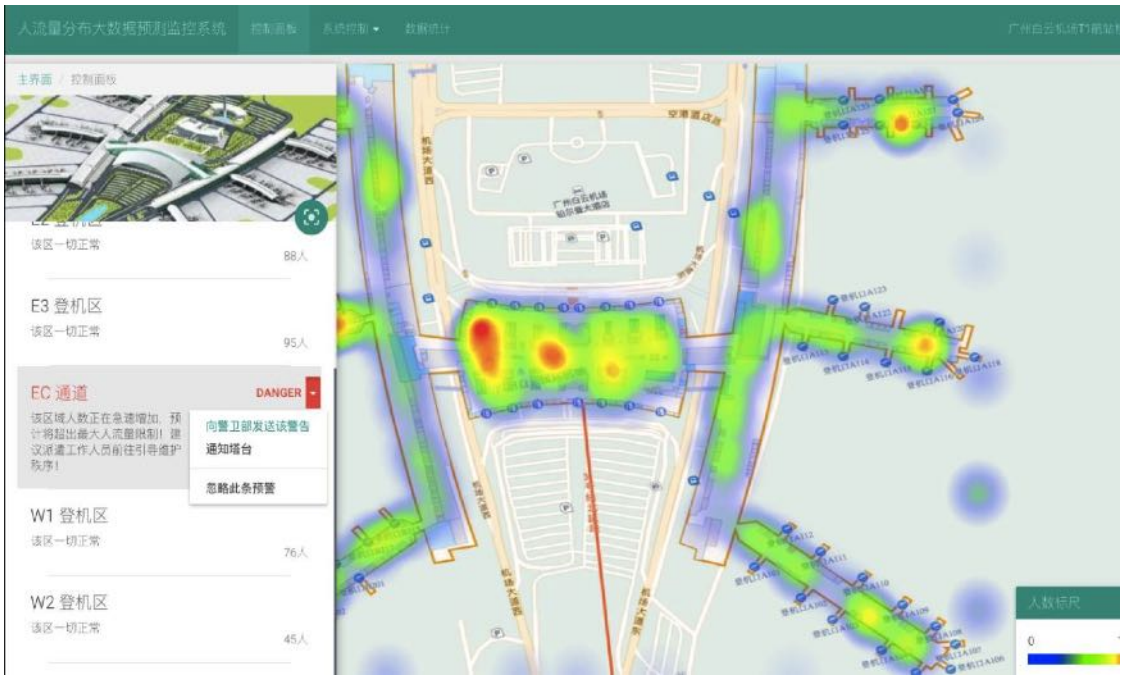


图 2.17 danger 警告措施处理

2.3.4.2.3. “系统控制”功能

“系统控制”选项包括启动系统、暂停系统、关闭系统，点击“启动系统”，系统就开始在界面显示实时人数热度图，并且可以拖动下方时间条，预测未来 10 分、20 分钟直至未来 2 天的人数分布：

图 2.18 系统控制功能图

点击“关闭系统”后，用户可以安全退出：

图 2.19 关闭系统注销界面

2.3.4.2.4. “数据统计”功能

选择某区域和预测时间后，点击主界面上方的“数据统计”，可以得出该区域从过去两小时到未来 2 天人数分析、预测折线图，未来某区域、某时刻（此处需用户自己拖动时间条设置预测时间）具体人数统计的柱状图和雷达图。

值得一提的是，本系统可以针对选择的优化措施给出相应的优化对比结果。

（1） 采取优化措施前：

显示采取优化措施前对应区域旅客数量，不同时刻的折线图和具体人数统计的柱状图和雷达图。

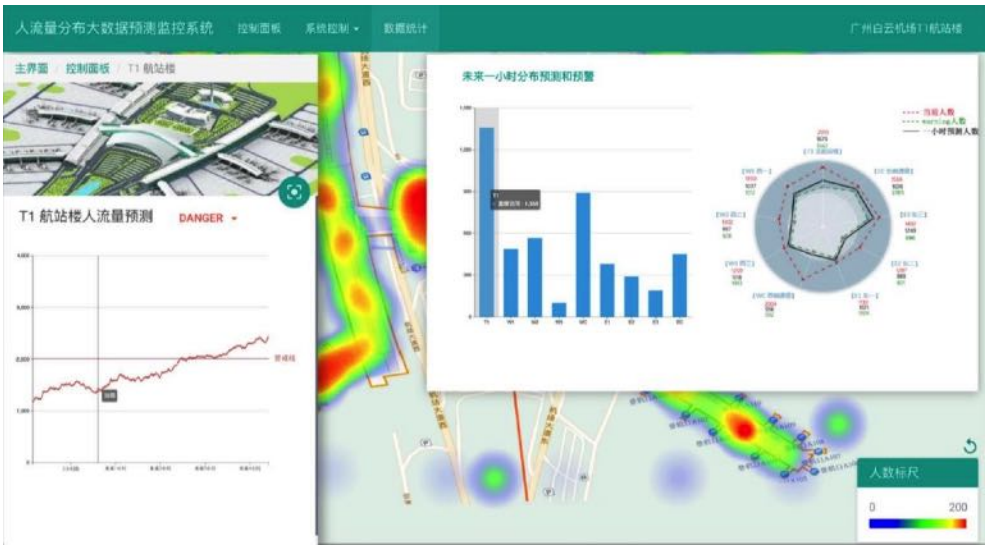


图 2.20 数据统计界面-采取优化措施前

（2） 采取优化措施中

显示采取优化措施过程中对应区域旅客数量，不同时刻的折线图和具体人数统计的柱状图和雷达图

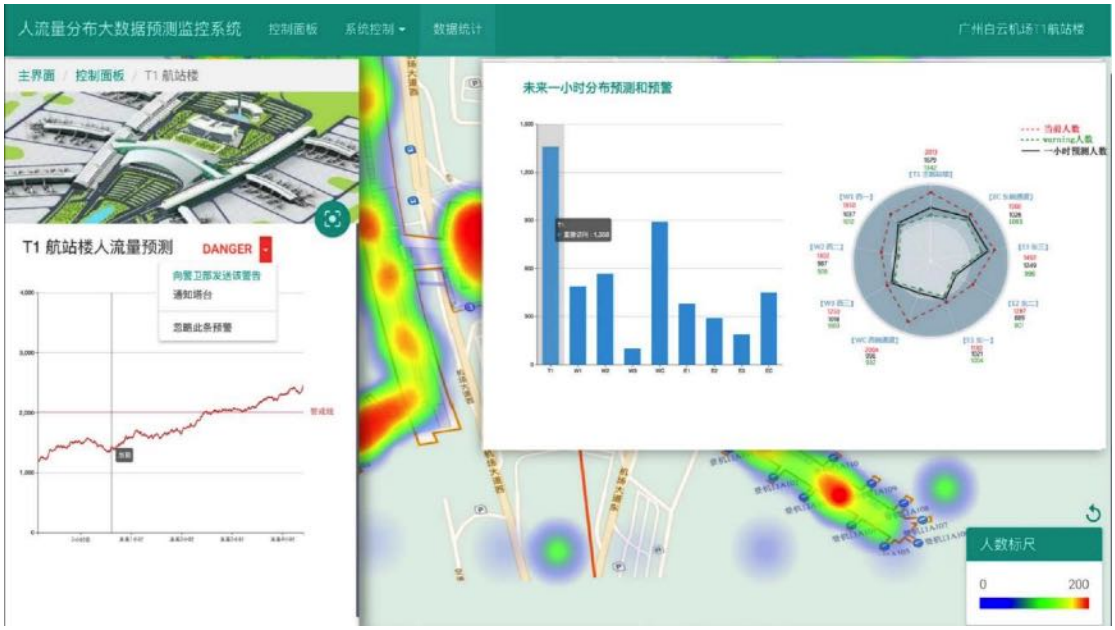


图 2.21 数据统计界面-采取优化措施中

（3） 采取优化措施后

显示采取优化措施过程中对应区域旅客数量，不同时刻的折线图和具体人数统计的柱状图和雷达图

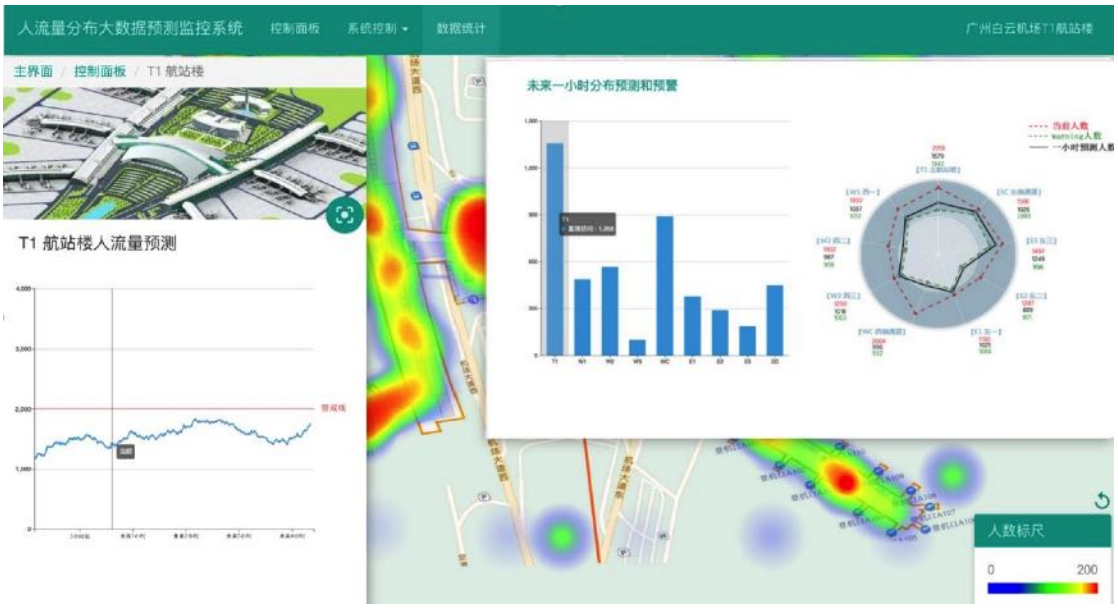


图 4.13 数据统计界面-采取优化措施后

2.4 升级与维护

2.4.1 产品升级

【数据方面】在数据收集方面，增加数据的收集来源和数据种类。采用本产品，根据新增的数据，分析不同种类的数据，得到更多的分析结果，以适应更多的应用领域。

【模型方面】应对不同的领域，模型特征有所不同。本产品可以应对多种领域，适当更改应用的特征，构建对分析有帮助的特征应用在不同场景，同时未来会增加新的模块实现新功能。

2.4.2 产品的维护与保障

由于产品涉及到服务器，为应对服务器可能会出现的硬盘受损等问题，我们设立备用服务器，并对实时采集到的信息镜像备份。对出现的紧急情况，安排紧急预案。

2.5 专利简介

2.5.1 基于 Spark 内存计算大数据平台的 OPTICS 点排序聚类方法

本发明提供一种基于 Spark 大数据平台的 OPTICS 聚类算法，涉及计算机信息获取和处理技术。本发明通过对并行数据结构划分，得到最优数据集划分并生成对应的 RDD，并行计算邻居样本数量和核心距离，对每个分区并行执行 OPTICS 算法得到每个分区的簇排序并持久化存储，通过簇排序给每个分区赋予簇后，通过合并分区，每个样本能够得到全局的簇号。利用 Spark 分布式并行技术，找到最优的划分结构，并行计算得到每个分区的簇排序。通过 OPTICS 的簇排序，用户可以从不同层次结构进行观察数据集的内在聚类结构。该方法能有效提高大规模数据的计算效率，更加符合数据量快速增长的实际应用场景，能够明显提高数据挖掘的效率，具有较好的实际应用价值且成本较低。

2.5.2 基于 Spark 大数据平台的三支决策不平衡数据过采样方法

主要的方法是重采样，增加小类样本的数目（过采样）或减少大类样本的数目（欠采样），目的是降低类别之间的不平衡性，有效地提高少数类的分类准确率和分类器的整体性能。

2.6 新服务研发

在本产品的基础上，我研发团队将继续对数据采集方式、采集种类进行扩展。并根据不同领域，开发相应的功能。

【针对机场】在预测系统成熟后，我团队将开发针对个别旅客的路线规划、消息推送等增值服务。并根据机场反映情况，对产品做细微调整，以达到最优化。

【针对扩展其他领域】团队会根据其领域的特性，增加不同的数据采集方式和种类，同时扩大数据采集范围，对其提供以未来人流预测为基础的，与其领域相关的功能，可包括：最优路线规划、智能空间（在某处得到所有与自身相关信息，智能航空楼）等。

第3章 市场分析

3.1 市场背景

中国民用航空局发布的公报显示，自上世纪 90 年代后，我国民航的主要运输指标一直有着平稳较快增长，民航运输需求不断增大。截止 2015 年，根据我国民用机场的客流量调查显示，全国民航运输机场完成旅客吞吐量 8.32 亿人次，同比去年增长 10.2%。

随着民航需求的逐渐增大，航空公司和机场所面临的压力也在增大。旅客服务水平一直是航空公司和机场关注的重点。对于航空公司来说，航空公司可以通过提高服务水平可在与其他航空公司竞争中提高竞争力；而对于一个城市来说，机场航站楼服务水平的好坏往往会影响旅客对于该城市的第一印象。机场也是一个重要的交通枢纽站，机场外围的交通运输设备的配置以及对于交通设施、执勤人员的调度都是城市交通不可或缺的重要环节。

民航旅客服务测评网站对自助值机等候时间、自助值机设备完好程度、Wi-Fi 上网服务满意度、机场航班信息通告满意程度等 14 个关键评测参数对国内机场进行了测评，测评结果显示，航空公司的空乘服务、客舱设施、机上餐饮等服务都较上一年有了明显提升，但值机和离港服务却一直处在评价不高的情况。2014 年，这类服务的评价反而下降了 3 个百分点。

值机和离港服务是旅离港流程中重要的一个环节。该环节评价较低的主要原因在于，旅客排队等待时间过长所致。过长的等待时间甚至会造成航站楼离港大厅拥堵、混乱等情况的发生。值机和旅客离港这个环节将直接影响到旅客的服务水平和机场的运营效率。这个环节之所以会存在如此大的问题，究其原因，不仅有值机柜台、安检通道资源数量不够等，更大的原因在于旅客数量随时间波动大，机场不能很好的预测客流量的时空分布，导致设施设备、人员调度不合理。

国内外学者对航站楼客流量预测方面的研究，主要集中在民航客运量、吞吐量上。目前，国内对机场内客流量的时空分布还仅停留在小数据范围内的研究，没有很好的利用历史数据进行数据建模，对于达到航站楼的各时段的客流量不能进行一个有效地把握。

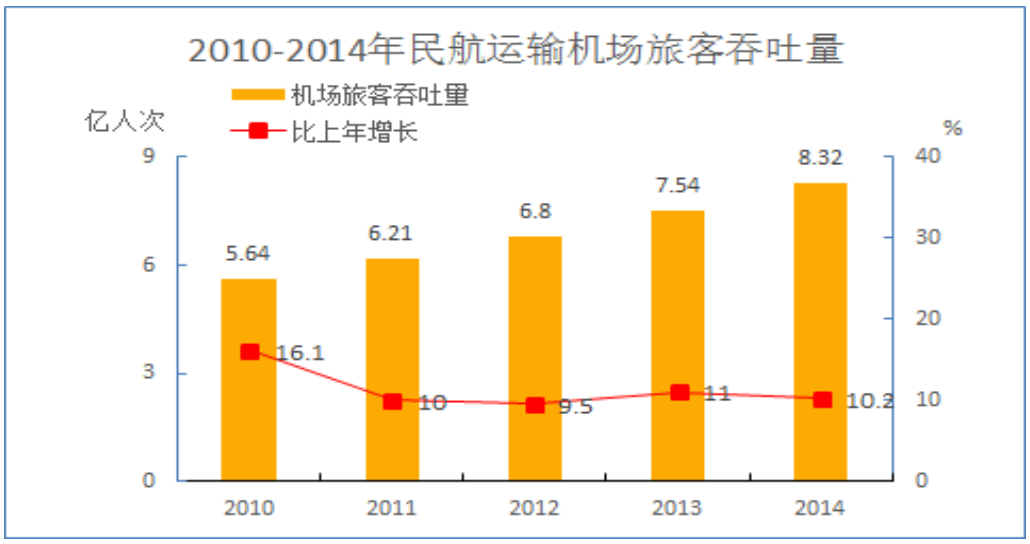


图 3.1 民航运输机场旅客吞吐量变化

3.2 消费者分析

产品的最终和最根本目标是为了服务消费者，因此，应当满足消费者的现有需求，激发消费者的潜在需求，只有充分满足了消费者的需求，才能保障市场的可持续发展。我们针对机场对于客流量预测的需求，从现有消费者、消费者动机、潜在消费者三个方面进行了分析。

3.2.2 现有消费者分析

【政府机关为主要消费者】一直以来，我国机场实行的半军事化的管理模式，政企合一、政事合一，机场行业多年实行的都是政府管制机制。机场不仅是一个重要的交通枢纽，连接城市的轨道交通和道路交通，也是一个城市极为重要的符号，是旅客对于一个城市的最初印象的评判。不管是航空公司或是航站楼的管理者都迫切希望机场能够有合理的调度安排和广阔的规划发展。

现在，我国经济强劲的增长与民航业的蓬勃发展带动着我国机场业进入了发展的“黄金时代”，与此同时，经济全球化和“航空时代”的来临，我国航空业也面临着巨大压力，不仅有国内竞争，国外的航空业的发展对于我国的各大机场拿过来说都是一个不小的压力。如何优化国内机场，合理配置资源，带给旅客更优质的服务以及让机场成为城市好的第一印象都是政府迫切需要解决的问题。机场以及机场外围的各项资源的安排调度都将是政府需要统筹规划的重点。

3.2.3 消费者动机

【提高机场竞争力】随着我国航空业的蓬勃发展，飞机成为越来越多人的出行选择。而用户体验将成为旅客评判与选择的重要因素之一。通过对于客流量的分布进行预测，提高机场的各项配置的合理性，调整外围交通带来便利，无疑可以提高旅客服务质量，让机场将在激烈的竞争中脱颖而出。

【附加属性和价值】随着国家对于航空业的规划与其自身的不断发展，机场的发展也不会只限于交通枢纽站，像是新加坡等国家的航站楼已经成为了一个大型商场，可以让旅客享受到更好、更优质的服务。

3.2.4 潜在消费者

对于机场航站楼的管理主要是政府机关单位。但现在国内的机场功能都基本只有单一的交通运输这一功能，机场占地面积大，客流量也多，机场未来的发展应该是多样化的，而不是单一功能，从现在机场内部有少数的商家就可以预见，机场外来的发展会更加商业化，而入住的商家和店铺都会对客流量的分布有着极高的需求，而这一需求就使得他们成为了产品的潜在消费者，且消费能力巨大。

机场作为交通枢纽站，对于机场周边的交通运输来说是不可或缺的一部分。机场周围的运力单位可以通过对机场客流量时空分布的实时预测，合理安排运力设施，缓解人流高峰期带来的拥堵现象的同时带动机场及其外围交通运输的发展。出于这样的需求，机场周围的运力单位也将成为本产品的潜在消费者。

本产品有强大的数据处理能力，适用于多个人流量大的公共场所，对于火车站、地铁站这样人流量流动大的地方本产品可以提供，数据分析以及设施管理，包括突发事件应急，行李追踪等服务；类似于商场、超级市场等人流量大的地方，本产品也可以针对不同的地方做出不同的管理方案，从而达到利益最大化。这样一来，地体站、火车站、商场、超级市场等人流量大的地方都适用于本产品，相关单位都会成为产品的潜在消费者。

3.3 STP 战略

我们根据市场现状，实行三步走的 STP 战略原则，具体分为市场细分、目标

市场、市场定位三个部分。通过市场细分选择目标客户，进而以此为根据确定目标市场，最后进行市场定位，以此来提升公司的核心竞争力。

3.3.2 市场细分

为有效把握市场脉搏，提高市场占有率，我们从客户类型、应用场景进行市场细分。针对产品的特性以及适用范围，我们将从两个方面对市场进行细分：

（1） 按照客户类型进行分类

现有的目标客户主要是国家政府机关，未来还会涉及到其他类型客户，扩大目标客户的范围，根据现有的目标客户进行市场细分情况如图所示：

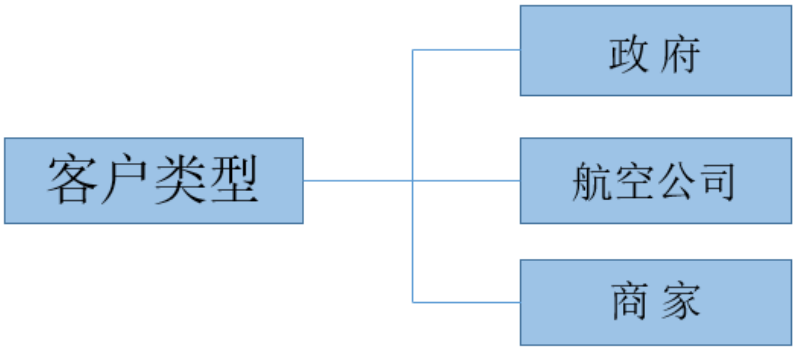


图 3.2 现有目标客户市场细分

（2） 按照应用场景进行分类：

本产品满足 IT 系统高性能、高可靠性、高安全性、低成本的需求，推动 IT 支撑系统集中化的实现，为业务系统增强大数据处理能力，打造互联网时代核心竞争能力。可适用于以下方面：

a) 机场、火车站、地铁站等人流量流动大的地方的数据分析以及设施管理，其中包括突发事件应急，行李追踪等；
b) 商场、超级市场等人流量大的地方的数据分析以及设施管理，针对不同的地方做出不同的管理方案，从而达到利益最大化；
c) 机场、物流等企业数据的分析与挖掘，做出最优的商务决策与发展战略；
d) 优化交通要道、车流量较大的道路上车辆的行车路线，减缓交通拥堵等问题。

3.3.3 目标市场

在细分市场的基础上，我团队在进行下一步的目标市场的选择。鉴于本产品是对人流量的时空分布进行实时预测且产品通过了广州白云机场数据的模拟测试，在此背景之下，本团队制定了不同时期的目标市场进入策略，最终面向所有细分市场提供产品与服务。

【初期发展扎根广州】在初期发展阶段，我们会先选择广州地区作为目标地区，因为后台数据来源于广州白云机场，模型对于白云机场会更有针对性，白云机场也是一个拥有巨大客流量的机场，全国排名前几名，有利于我们快速扎根目标市场，打下基础。

【区域拓展】在导入期，我们决定采取区域拓展的战略，将目标市场的区域以广州白云机场为中心，向内向外拓展，进而拓展到国内的大小型机场。

在成熟其和衰退期，我们的目标市场就是全国的所有机场，并且会根据产品应用效果的不同和战略期的不同，制定市场战略，逐步应用到不同的场景中，占领目标市场。

阶段	目标市场	市场特征
培育期	广州白云机场	白云机场客流量巨大，离市区相对较近。有较为真实数据模拟测试。
占领期	以机场为基础辐射周边运力场所	白云机场已经应用成功，机场周围需要进行运力调配。
拓展期	全国各大小型机场	机场客流量相当庞大，全国的机场都需要优化配置，提高服务质量
领航期	其他各客流量大的公共场所	人流量时空分布的实时预测适用于各拥有巨大人流量的场所

表 3.1 目标市场

3.3.4 市场定位

服务定位：坚持为目标客户提供高效、合理、安全的资源调配方案。

竞争定位：依靠差异化的竞争策略成为客流量时空分布实时预测的领跑者。

3.4 竞争分析

国内对于用算法和模型来预测人流量的时空分布应用与开发起步较晚，无论从实地应用还是从发挥效益来看，预测数据的应用价值没有得到充分的挖掘，所以国内相关产品的开发与应用还有很大前景。因此，本团队在综合各方面的考虑之后，对于各个战略时期进行了分析，得出各战略时期的竞争威胁，并制定相应的竞争策略，鉴于各期的竞争威胁侧重点不同，本团队详细的分析各战略期的竞争概况。

3.4.1 培育期竞争分析

【现有预测系统威胁】在培育期，关于机场的相关客流量的预测系统的竞争威胁来自于现有的预测系统，比如说对于机场客流量的吞吐量的月度、季度、年度预测，小范围内的人流量的分布预测。

【本产品优势显著】团队所提供的产品是基于大数据平台和数据挖掘技术的模型，有效地解决了只能小范围、长时间的不精准预测模式，克服了现有系统存在的无法精准到分钟、不能出解决方案等问题。不仅给机场带来了巨大的经济效益还给国家带来了社会效益，同时对于资源的调配也控制了成本，提高了效率。我们的产品作为市场的补缺者，通过了广州白云机场的真实数据的模拟测试，因此，现有的预测系统对我们的项目系统威胁不大。

3.4.2 占领期竞争分析

【潜在竞争者威胁】在占领期，我产品面临的竞争威胁主要来自于现有的竞争者、潜在竞争者的威胁。市场目标具有面向社会、广泛招标的特点，在交易中往往处于主动地位，但我们的产品在精准度和应用场景方面都优于现有的预测系

统。对于潜在的竞争者来说，本产品以先进的技术进入了市场，取得了政府机关较高的忠实度，形成了良好的口碑，且技术力量雄厚，服务不断推陈出新，保持了较高的市场竞争力。

【国外竞争者威胁】与此同时，在占领期，国外的竞争者也对我们的产品构成了重要的威胁。我国客流量的时空分布预测应用本就起步较晚，随着对外交流的密切，国外的一些较为成熟的技术进入我国市场的可能性将增加，但国外的机场大多交由市场调控并非是政府管制，跟我国国情有很大出入，其次我们产品已通过了国内大型机场的试点模拟测试，取得了市场先机，并且有技术优势，具有先发优势。

通过对国外竞争者和潜在竞争者的分析可知，本产品在各个方面均具有明显的竞争优势。在占领期，我们也将采取目标集聚的竞争战略，通过公关保持与政府机关的长期合作关系。因此，**国外竞争者和潜在竞争者对我们的威胁均不大。**

3.4.3 拓展期竞争分析

【潜在竞争者为主要威胁】在拓展期，本团队面临的竞争威胁来源于现有竞争者、潜在竞争者的竞争威胁。现经过前两个战略期的市场开拓，政府机关及供应商的价格体系基本形成，因此，议价威胁降低。此阶段，我们主要的竞争威胁来自潜在竞争者。

随着我们在区域市场的扩张以及服务行业的扩大，获得的利润也将逐年增长。受利益的驱动，更多的企业将进入相关行业，与我们抢占市场份额。

【大力拓展国内市场且关注国外市场】在拓展期，我们已有了较高的市场占有率。我们主要通过市场改良、产品改良的产品策略，以合理的促销策略等方式大力拓展国内市场，不仅要贯穿到全国各个机场、各大型公共场所，还要十分关注国外市场，开辟一条通向海外市场道路。因此，**潜在竞争者对我们的竞争威胁不大。**

3.4.4 领航期竞争分析

【替代品威胁为主要竞争威胁】在领航期，我们面临的竞争威胁来源于现有竞争者、潜在竞争者、替代品的竞争威胁。该时期我们主要竞争威胁来自于替代

品的威胁。

随着大数据行业的逐步成熟，可观的行业利润、成熟的研发技术、政策支持及行业的发展趋势必然会吸引更多的企业进入相关服务行业，给我们带来一定的威胁。

【采取多种方式保持市场领导者的地位】客流量实时预测服务对相关技术要求高，而我们将及时对系统进行维护与升级，保持行业领先水平。同时，我们将坚持目标集聚辅以差异化的竞争战略，保持市场领导者的地位，采取开发新产品、品牌延伸的产品策略，因此，替代品对我们威胁较小。

3.4.5 竞争力概况

本产品创新地将大数据预测应用在机场人流分布这一全新的场景中，根据机场的人流分布，提供优化建议，使航站楼内的各类灯光电梯设施设备、值机柜台、商铺、广告位等安排更为合理，能更加精准、高效地调度这些资源和安排服务人员，减少机场该部分费用的同时也保障了机场安全。

在系统功能上，除了可对机场人流分布情况进行实时监控，还可以以 10 分钟为间隔，连续动态且快速地预测未来 10 分钟、20 分钟、30 分钟等的人流分布，并可让用户自由设置时间间隔。特别的，产品可针对不同的人流分布情况给予不同的合理化建议，并将优化前后的情况清晰明了地展现在用户眼前。

产品模型以白云机场为落地案例，设计相应的人流分布预测和预警功能，可在全国机场做推广和复制，除此之外也可用于火车站、商圈等场景，前景广阔

3.5 未来市场展望

就人流量时空分布的实时预测而言，我国国内的市场极为广大。随着我国大数据等相关技术的飞速发展，这一技术将会在各行各业广泛运用。对此，我们也做出了一些展望。

（1） 扩展模型功能模块：开发并实现模型更多的功能，提高系统的数据处理、分析能力，展现更多的预测模块数据，从而进一步提升预测数据的质量，充分保证分析结果的真实性和准确性以及预测结果的可靠性，也给使用者提供更多的数据分析；

（2） 本作品以白云机场为落地案例，可在全国机场做推广和复制，为其他机场提供优化建议，使航站楼内的各类灯光电梯设施设备、值机柜台、商铺、广告位等安排更为合理，能更加精准、高效地调度这些资源和安排服务人员，减少机场该部分费用，增强机场安全；

（3） 推广到其他公共服务区：将数据搜集方式扩大化，兼容更多数据采集方式。将完善后的模型算法推广到类似地铁站、商圈等人流量大的区域，按照需求细化产品功能；

（4） 提高系统处理能力：从数据处理和预测数据两方面提高系统对于数据提纯的性能，支持处理更大量、更多维的数据；

（5） 改进和优化：作品功能还不够人性化和智能化，广泛听取更多使用者的意见，调整模块，增加人性化特色功能。降低数据挖据发开难度，降低维护成本。

（6） 提高系统存储能力：从数据处理效率和数据存储规模两方面提高系统存储性能，支持更大量级的数据存储能力。能够有效地保存之前几天已经发生的数据以及预测到的几天之内的数据，从而便于用户查询以及机场的管理。

相信通过我们的努力与思考，对产品做出进一步的研发与改进，使其能够为社会创造更多社会价值和财富价值，为人民带来更多便利，提高人们出行的舒适度和幸福感。

第4章 作品的推广及应用

4.1 科学性及先进行分析

4.1.1 科学性分析

在人类社会发展进程中，人们通过积累经验预测自然现象、把握自然规律来改善自身生存状况的行为从未停止。不管人们以怎样的方式观察、采集这些信息，最终都以数据形式存储。以前人们在大量数据中寻找规律，现在人们用规律预测未来。数据已然成为与自然资源、人力资源一样重要的战略资源。大数据时代的到来，促使人们采用超越前人的大量数据，以大数据技术为方法，为社会、经济和科技的发展带来了重要机遇，也对数据的存取、传输、计算提出了新的挑战。

介于大数据发展的必然性，各类大数据团体相继成立。深圳大数据产学研联盟于 2013 年成立，2012 年美国启动“Big Date”计划，并发布了《大数据研究与发展倡议》，2014 年中国电子技术标准化研究院制定的《大数据标准化白皮书》中指出，我国大数据工作应重点“加强大数据核心技术研究、推动开放数据集建设、创新大数据应用模式”。另外，大数据改变着企业的运营方式，从掌握已有客户到发展潜在客户，洞察客户的需求，在经济、社交等方面给与客户更好的服务成为其竞争的关键。

本作品借由预测未来 2 天内的旅客人流流量，调配机场设施，已给旅客更舒适的服务，并防止意外事件的发生。在机场设备调用上，本作品可以延伸到机场配置的大巴车、连接机场的轻轨线，缓解这类交通要道的拥堵情况。在界面上，研究人员可以扩展、完善更多功能，提供更多便利。不仅是对于机场管理人员，对于准备在机场候机的旅客都适用。从搜集数据的方式就可以发现，只要保证有实时的 Wi-Fi 点，就可扩展到更多大型公共设施领域使用，以确保设施的有效性和游客的安全问题。甚至在本次模型的基础上，可以设计更多的数据采集方式，得到更多的领域数据[25]，有效检测、预测、管理相关地区的人流量对社会安定发展有极大的意义。

4.1.2 先进性分析

使机场调度人员，安放设施更加合理高效，且智能化。 除此之外还可以根据实时情况提供优化建议，并显示出优化前后的对比。

智能化：根据机场当天的实际情况以及需求预测相应的几天的情况，并做出相应的改动。

高效：短时间内可能有大量的数据需要进行处理，对数据流的处理流程做到了简单、快捷，并且能够及时预测。

系统智能化地分析数据，显示实时情况，当某区域人流密度过大，能提供优化建议，减缓这种情况，并显示出优化前后的对比。

4.2 适用范围与推广前景

4.2.1 适用范围

机场拥有巨大的旅客吞吐量，与巨大的人员流动，相对应的则是巨大的服务压力。本作品提供的大数据分析模块化基础平台不仅帮助机场预测未来几天的人流量分布，而且致力于在安防、安检、突发事件应急、值机、行李追踪等机场服务方面能够帮助到机场及时了解情况，并据此提前调配人力物力，更好的为旅客服务。同时，也能够使机场准确把握当前最新发展动向，及早发现行业市场的空白点、机会点、增长点和盈利点，前瞻性地把握该行业未被满足的市场需求和趋势，形成企业良好的可持续发展优势，有效规避该行业投资风险，更有效率地巩固或者拓展相应的战略性目标市场，牢牢把握行业竞争的主动权。

本作品满足 IT 系统高性能、高可靠性、高安全性、低成本的需求[26]，推动 IT 支撑系统集中化的实现，为业务系统增强大数据处理能力，打造互联网时代核心竞争能力。可适用于以下方面：

（1） 机场、火车站、地铁站等人流量流动大的地方的数据分析以及设施管理，其中包括突发事件应急，行李追踪等；

（2） 商场、超级市场等人流量大的地方的数据分析以及设施管理，针对不同的地方做出不同的管理方案，从而达到利益最大化；

（3） 机场、物流等企业数据的分析与挖掘，做出最优的商务决策与发展战略；

（4） 优化交通要道、车流量较大的道路上车辆的行车路线，减缓交通拥堵等问题。

4.2.2 推广前景

国内对于用算法和模型来预测人流量的时空分布应用与开发起步较晚，无论从实地应用还是从发挥效益来看，都与国外有一定的差距，预测数据的应用价值没有得到充分的挖掘[27]。现在国内把相关技术应用于设施设备建设、人力资源管理、调度系统的产品很少，所以国内相关产品的开发与应用有很大前景。

同时机场作为由政府主导的特殊企业，兼具了商业性和公益性。当二者发生冲突时，加上一些不可抗力因素导致的机场停班等状况，须由政府提供补贴，使其能够发展下去，长此以往，机场经济入不敷出。2015 年中国民航支线航空论坛指出，支线航空的补贴增长速度大于全民航的游客增幅，其中对机场补贴达 12.11 亿元，同比增长 12.29%。所以改善机场的经营势在必行。本次的例子白云机场很早就提出了《论机场航空性业务的收入和发展》一文，开展非航空性业务是机场向企业化经营的必然要，也是开展多种经营、实现规模效益的必经之路。在现有资源的基础上，制定以民用航空地面运输保障为中心，建立集商业、科技、旅游、餐饮等多层次，实现多元化、一体化、国际化的经营战略。

所以在基于良好的服务质量的同时，利用大数据技术，本作品也可将信息采集定位在旅客的兴趣爱好和相应需求，在此基础上，通过模型、算法分析，挖掘潜在的旅客信息，分析了解个体客户的实际需求，结合机场的各种资源在正确的时间，以正确的方式向旅客提供最优的咨询，引导旅客消费，既扩大旅客的满意度，也使机场的经营模式多样化。让航站楼不仅作为空港存在，更使其成为旅客在出行中的一处消费、旅游场所，以提高其经济效益。

4.3 社会效益与经济效益

（1） 本作品采用了基于 Spark 内存计算大数据平台的数据挖掘技术，实现了科技实用化，以实际机场为作用点，增强科技的实用性，鼓励更多专业人才应用

先进科技技术，建设国家。

（2） 本作品顺应大数据相关行业的发展，以机场为基础，根据 Wi-Fi 点的坐标和人数，进行聚类可以快速得到人群聚集在哪些地方，可提高机场的人员设施配置合理度，减少了机场因人员设施的调配安排而产生的开销，除此之外，本系统也具有对意外事件的预警能力，一定程度保证了机场的安全，同时也提高了客户的舒适度。

（3） 本作品应用范围广泛，可变性强，除机场外，其他大型公共设施点也可使用，类似于广场、火车站，轻轨站等，都可用做人员设备调配的依据，减少意外事件的发生，也可对潜在危险进行防范，有助于社会安定，提高人民的幸福感；

（4） 此外，本作品适用的场景多样，除机场工作人员外，还可作为商家的参考数据，帮助商家布置相应的贩卖设施，筛选合适的营销地点，制定合适的营销策略，极具商业价值，也可为交通繁忙的地段预测未来车流并提供合理化建议，除此之外还可以为现代物流企业提供中小仓库建仓地点的建议，为物流企业节省不必要的开支等等.

第5章 商业模式

5.1 用户需求分析

2015 年中国民航支线航空论坛指出，支线航空的补贴增长速度大于全民航的游客增幅，其中对机场补贴达 12.11 亿元，同比增长 12.29%。空域资源配置不合理、基础设施发展较慢、专业人才不足、企业竞争力不强、管理体制有待理顺等制约了民航业的可持续发展，所以改善机场的经营势在必行。

首先，民航机场作为航空运输的基础设施，是综合交通运输体系的重要组成部分。人们对航空领域的需求量增加，使得各地都兴建机场，直接导致机场行业的竞争越来越激烈。机场服务质量和服务管理水平作为衡量机场优劣的两大基本要素，在客户对机场要求日益提高的今天，对机场的发展起到决定性的作用。如何不断提高服务质量和服务管理水平，满足旅客日益增长的需求，提升机场企业的核心竞争力，已经是民航机场广泛关注的重要课题。

所以从机场管理角度出发，如何制定有效的管理方法，以提高服务质量，优化机场设施？众所周知，由于机场的特殊性，机场最大的隐患就是人流量的问题。庞大的人流量，影响了机场各种配套设施的使用。从旅客进入机场的那一刻起，办理登记手续，托运行李，安检，候机，甚至领取行李，转其他交通工具，旅客随时都处在排队状态。其直接导致了大量旅客滞留在机场范围，占用大部分的机场设施。由于员工和设施有限，机场不能够为所有客户提供其所能及的服务，服务质量降低，其直接后果是机场的口碑下降，竞争力下降。

其次，由于机场环境的特殊性，考虑到机场建设面积、居民生活环境等一系列因素，我国机场大多建设在远离市区、居民区的地方，所以，“机场周围交通不便”也是近年来，旅客在航空行业反映的较为突出的问题之一。

综上所述，机场无论是作为空港还是一个重要的交通枢纽，需要给来客提供一个舒适、安全的环境。再者，机场作为运输科技的代表，可以由己推广到更多领域，有模范的带头作用。

5.2 目标用户群体

（1） 在培育期，我们以机场为主，相连的交通枢纽为辅。机场内人员流动量大，相对应的则是巨大的服务压力。本作品提供的大数据分析模块化基础平台不仅帮助机场预测未来几天的人流量分布，而且致力于在安防、安检、突发事件应急、值机、行李追踪等机场服务方面能够帮助到机场及时了解情况，并据此提前调配人力物力，更好的为旅客服务。同时，也能够使机场准确把握当前最新发展动向，及早发现行业市场的空白点、机会点、增长点和盈利点，前瞻性地把握该行业未被满足的市场需求和趋势，形成企业良好的可持续发展优势，有效规避该行业投资风险，更有效率地巩固或者拓展相应的战略性目标市场，牢牢把握行业竞争的主动权。

（2） 拓展期群体。在产品扩展阶段，通过升级模型，目标用户群体也有扩大。

a) 对于机场，在基于良好的服务质量的同时，利用大数据技术，本作品也可将信息采集定位在旅客的兴趣爱好和相应需求，在此基础上，通过模型、算法分析，挖掘潜在的旅客信息，分析了解个体客户的实际需求，结合机场的各种资源在正确的时间，以正确的方式向旅客提供最优的咨询，引导旅客消费，既扩大旅客的满意度，也使机场的经营模式多样化。让航站楼不仅作为空港存在，更使其成为旅客在出行中的一处消费、旅游场所，以提高其经济效益。

b) 对于其他公共设施场所，如广场、商圈等，人流吞吐量大的地方，运用本产品都可用做人员设备调配的依据，减少意外事件的发生。

c) 对于个人用户，由于某些原因需要提前得知相应地区的人流状况，以便提前安排行程。本产品可以提供，实时预测的人流分布图，帮助商家布置相应的贩卖设施，筛选合适的营销地点，制定合适的营销策略，极具商业价值。

5.3 商业模式选择

5.3.1 培育期

【用户无偿使用】在培育期，这种模式主要针对机场。机场可以免费试用我们的产品，获取预测的人流信息，以实现设备优化调整。

且会对相应的机场状况提出实时的意见，使用户能感受到使用产品前后的不同。

5.3.2 占领期

【试用期+用户有偿购买服务】本时期中，主要针对机场及周边的运力单位，提供三个月的试用期。在使用期结束后，用户需支付相关费用，获取客流信息，实现联合运力调配，共同获利。

5.3.3 扩展期和领航期

【产品与技术混合销售】在本时期，产品的覆盖范围扩大，除了有偿为一些单位提供产品，为其搭建服务器外，同时向其他领域的用户，提供技术服务，实现以此模型为基础的，用于其他信息预测的功能。

5.4 盈利模式

5.4.1 购买产品服务收入

【产品售后】在产品推广后，团队可建立相关的信息库，对只购买产品服务的用户，提供信息保存等售后服务，从而获得利益。

5.4.2 购买技术服务收入

【产品更新】在产品拥有固定大规模的用户群体后，通过团队的研发，本产品功能将会不断完善，根据不同的情况，又有不同程度的更新。针对购买技术的

用户，我们将根据其实际情况，给出类似于系统维护、更新技术等技术性较高的产品更新服务。在其遇到技术障碍时，也可提供有偿的技术服务。

5.4.3 购买产品收入

【销售完整产品】主要通过在不同的领域推销本产品来获利。本产品以机场为试运点，向其他运力单位，商圈的地方发展，再涉足其他领域，推动整个社会的科技发展。

5.5 客户关系

本团队致力于为客户提供预测服务，并为客户提供信息保存等需求。其预测面可由前期的人流预测扩展，提供针对不同领域信息服务，从而建立与客户的联系，相互促进，发展科技、经济，维护社会的和谐、稳定。提高人民的幸福指数。

5.6 合作伙伴

目前本团队已与重庆邮电大学大数据研究院合作，共同研发新型功能，扩展产品辐射范围。在此基础上，团队也将通过盈利分成方式，协助我方产品在其他用户群体上的使用。

第6章 营销模式

6.1 营销理念

我团队产品将秉承“适应市场，满足市场，创造市场”的营销理念，结合市场分析，充分利用地缘关系、亲缘关系、业缘关系等资源，力争广州白云机场的支持，并在广州、重庆等省市机场开展试点工作，打造示范点；将示范点成果向中国民用航空局汇报，获得中国民用航空局的大力支持；进一步以示范点为基础，推广大数据服务，塑造良好的品牌形象，从而带动全国的市场。

6.2 营销影响因素分析

表 6.1 营销影响因素分析

影响因素	自身因素	产品与服务因素	市场因素
因素体现	初期的网络营销需要时间进行覆盖 缺乏较为专业的营销团队	产品缺乏知名度和影响力	机场大数据客流量分析产品市场处于发展初期 缺乏厂商提供较为完整的服务解决方案
应对策略	提高产品的知名度和影响力以吸引客户 打造/聘用专业的营销团队用更专业的营销手段运营和推广	强调产品的功能的可用性以及初期的无偿使用 突出产品针对问题能高效智能的解决方案	深入挖掘市场 提供差异化的产品与服务

6.3 4P 理论营销策略

【客流量大数据分析市场前景广阔，营销策略由市场决定】客流量大数据分析行业还处于高速发展阶段，市场还远远没有出现饱和状况，本团队产品客流量大数据分析系统还拥有广阔的市场前景。为了全面打开市场，迅速占领市场份额，本团队针对系统产品在推广时面临的一系列问题，制定了详细的营销策略来应对瞬息万变的市场与不断变化的消费者需求。

【系统简洁，可视化强】当前社会已经存在了部分大数据分析企业，但是他们采用的方式过程繁琐且消耗时间较长，在结论上也没有提出相关的解决方案，

以至于无法给客户提供直观、简洁的分析结果。本团队自主研发的机场客流量大数据分析系统，数据获取方面丰富，操作简单，能做到分析结果可视化强，使之具有更高的工作效率，为企业合理解决机场存在的相关问题提供了强有力的参考依据。因此，本团队在机场客流量大数据分析这一方面具有得天独厚的优势。针对本团队的这一巨大优势，并且结合不同时期的市场特征，我们根据经典的 4P 理论，制定了详细的营销策略，以便快速赢得消费者的青睐。

6.3.1 产品策略

（1） 产品介绍：

【首创PC端与移动端双端操作】本系统，首创PC端和移动端在线同时操作，客户在 pc 端看到可能出现的特殊情况，可及时派人处理危机情况，工作人员也可以用移动端及时看到机场相关信息。

【可视化操作简单】整个机场数据客流量的获取、验证、预测到提出解决方案各环节操作流程简单，为用户提供简洁明了的使用界面。系统无需用户进行复杂的操作，只需打开本系统客户端，按指定流程操作即对机场客流量有直观的感受，无任何计算机技术背景的用户均可在短期内熟练操作。此外，本系统细化产品应用模块，根据用户需求不同，机场客流量大数据系统推出不同的增值模块，客户可根据个人需要单独定制购买，全面提供客流量预测与解决方案。

（2） 产品差异化策略：

【突出实时分析的优势】重点突出本产品可以进行线上实时分析越策的优势。通过前期调研，发现客流量大数据分析实时预测仍然没有实现，许多客户苦恼于无法有效解决人流量过大、安全事件频出问题。本系统是第一个实现该技术的产品。我们将自己的产品定位为“为企业提出实时解决方案”，通过充分强调这一特性，强调我们与大数据分析系统的差异化和先进性，戳中用户痛点。

（3） 改良策略：

【产品改良策略】一是升级产品，不断适应市场需求，适应客户的定制要求；二是提高机场客流量大数据分析系统的直观性、可靠性、安全性等；三是改良机场客流量大数据分析系统的用户体验，提高界面可视性，减少操作难度等。四是附加产品改良，向消费者提供良好服务、优惠条件、技术咨询、质量保证、消费

指导等。系统改良后再度投放市场，在巩固原有领域的基础上，拓展更新更多业务，争取更多的潜在用户，不断巩固市场领导者的地位。

【市场改良策略】不同的市场产品所处的时期不一样，有些已经到了成熟期，有些市场则可能刚处于成长期，因此改变策略，把产品重点向成长期市场推广。在信息技术相对落后的中西部地区，甚至是国外市场都可以考虑开拓，力求延长机场客流量大数据分析系统的成熟期。可以通过宣传等手段，使客户提高对于产品的实用频率，也可增加销售量，也可以达到延长成熟期的目的。

6.3.2 价格策略

【撇脂定价】互联网数据取证服务市场潜力巨大、尚未开拓、用户价格敏感度较低，并且本团队创业团队拥有自主研发的专利技术。在前期，使用本产品的主要使用场景是机场，个人消费者较少，因此，为了迅速收回成本、降低财务风险、塑造高端的品牌形象，决定采取撇脂定价策略，赚取较高的利润，为后续开拓市场做准备，在后期随着规模的扩大，使用本产品的消费客户的逐渐复杂化以及消费水平的差异化，再逐渐降低价格。

【导向定价】一方面根据市场竞争力的大小，对产品的价格进行调整，在市场竞争力较小时，可以制定比较高昂的价格，获取高额的利润，以求快速回收成本。另一方面根据核心客户的消费能力来进行定价，将客户定位为高端消费者时，可以制定高昂的价格来回收成本，将客户定位为大众消费者时则要调低价格，以数量的优势来弥补价格上的缺陷。团队要结合两方面的因素，制定最为合理的价格，以降低销售风险。

【免费试点】系统研发之初，为完善产品，打响知名度，特与广州白云机场开展合作，上述机场为本产品提供试点，而本团队免费为其提供一段时间的客流量预测以及解决方案。双方互利共赢，以便后期开展更深层次的合作。

6.3.3 渠道策略

【直销策略】新兴产品很难一开始就打开市场，先可以通过立足白云机场进行试点营销，发展机场周边并且走向全国各地机场市场为导向逐渐打开各大市场。初期主要发展白云机场，中期重点发展机场以及地区周边，长期开拓全国机场以

及人流量较多的市场。产品在打响了一定知名度，已经有相对规范成熟的销售渠道后，逐渐在机场周边地区开展大范围直销业务，大举切入以及全面覆盖全国各大航空企业单位等市场。

6.3.4 促销策略

（1） 会议营销

【研讨会】召开“航空领域大数据发展和建设”等主题的学术研讨会，邀请重庆市相关政府官员、航空公司等企业代表、专家学者参与交流讨论，借以推荐该产品。及时参与各类其他相关主题的专题报告会，向相关机构汇报介绍我们产品的总体方案和规划设想以及初期的建设情况。还可通过高端会议营销，提高产品公信力，及时参与其他各类相关会议。

【展示会】当我们的产品 A-guardian 要想在一个新地区、新市场进行推广发展时，由于客户对于产品的认知度不高，产品影响力较小。因此我们通过参加或主办互联网展会的方式，邀请专家学者、行业代表、专业观众共襄盛会。每年的各类互联网大会都有一大批高新技术产品被人们所熟知。我们通过这些展会，介绍我们的产品并进行推介。除了线下展会外，还可建立网络虚拟展会，用户可直接通过网页浏览我们的产品，获得更详尽的介绍。

（2） 公关营销

【公关交流】本团队将通过产品服务热线、企业专访、座谈会等方式做好公关工作，加强与各单位，各方面的交流合作，建立高效的信息交流反馈机制。及时调研，了解客户需求，提供更加个性化的定制服务。公共关系的构建不仅可以为销售提供保障而且有助于提升品牌形象，我们将重点建设以下公共关系网络，与政府机关合作关系的构建；与运营商、代理商等合作关系的构建；推行品牌营销战略，赋予公关实务以新的载体；转变竞争方式，建立以“多赢”为中心的公关理念。

（3） 网络营销

当今互联网用户大量增长，因此任何营销都离不开互联网，借助互联网营销，具有成本低、环节少、跨时空、方式新、市场冲击性强等特点。因此，除了在实体市场上的各类营销，团队将在本阶段同时采用多种网络营销方式。

企业网站	搜索引擎	互动营销
介绍、宣传企业产品的网站	Google、百度、火狐等主流搜索平台	微信公众号、微博、自媒体的运营

图 6.1 互联网营销方式

（4） 人员推销

【实时掌握顾客户信息】成立专门的运营团队，负责产品的推广工作。在推销过程中，买卖双方当面洽谈，易于形成一种直接而友好的相互关系。通过交谈和观察，可以掌握客户的需求动机，有针对性地为客户介绍针对他们的具体需求，我们的产品所展现的特点和功能以及所能够提供的服务和带来的便利。根据客户的态度和特点，有针对性地采取必要的协调行动，反馈给研发团队以满足客户需要，抓住有利时机促成交易。还可以及时发现问题，进行解释，解除顾客疑虑，使之产生信任感。吸引公司与相关企业保持密切关系，培养用户忠诚度，提高收益率。

（5） 专家推介

本产品成功研发后，邀请了相关领域的数十位专家、学者对本产品进行了深入的了解体验，取得一致好评。

6.4 “一对一”营销策略

“一对一”营销为我团队和客户间的互动沟通提供具有针对性的个性化方案，目标是提高短期商业推广活动以及终身客户关系的投资回报率，最终目标是提升产品的整体客户忠诚度，使客户满意和客户价值最大化。“机场流量大数据分析系统”是一个服务性科技产品，其目的是更好的服务机场，带动周边产业促进区域经济发展。我团队客观的分析客户需求，并针对客户需求，设定了“一对一”营销策略。

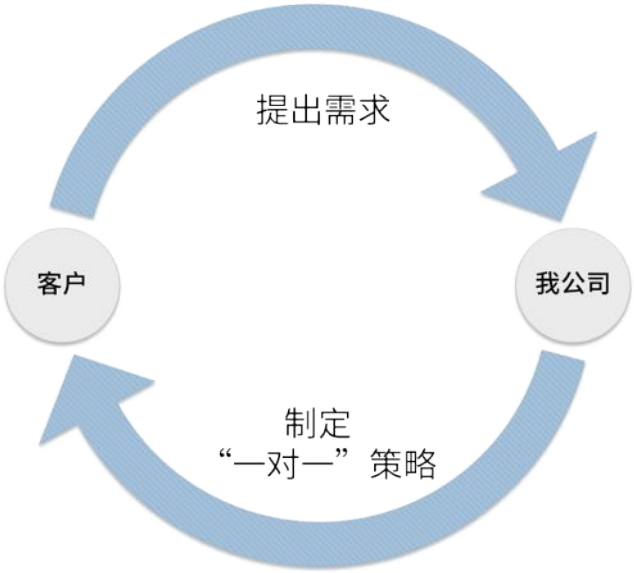


图 6.2 “一对一”营销策略

6.4.1 客户决策策略

（1） 成本因素

客户在选择产品时会优先考虑他购买的产品所产生的价值是否足够抵消他购买产品时所花费的成本，或者是否在他所能承受的范围内，这便是客户选择产品的成本因素。例如本团队产品“机场客流量数据分析系统”，客户在购买之前便会衡量其带来的作用，是否足以抵消购买它所产生的成本。

（2） 便捷因素

客户购买产品时，往往会考虑购买产品的途径和产品的操作难易，太过繁琐的购买途径，往往会导致一部分客户放弃购买产品，太过繁琐的操作流程也会使客户对产品的使用丧失兴趣，购买的欲望急速下降。本产品能够进行在线实时预测机场流量并在前端显示，且经过团队培训后操作人员可掌握系统使用方法，这也是足够影响客户决策的因素之一。

（3） 交流因素

在初期，客户对于产品的了解还处于比较模糊的阶段，对于产品没有一个清晰的认识，特别是像本团队推出的“机场客流量大数据分析”这类的高科技产品，客户的知识有限，对于产品的了解存在局限性，客户需要与企业产生有效的沟通交流，才能对产品有所了解，同时产生购买的欲望。现在市场上无相关有效产品，因此相关客户也没有学习对比的经验，如果没有进行有效的交流，这可能导致客户对于本产品存在认识的偏差，客户购买力的下降。

6.4.2 企业应对策略

（1） 产品策略

由于市场上同类产品较少，团队主要将通过机场突发事故（如特殊天气造成的延误或航班取消）的实例分析，得出产品的优势，以产品的优势为切入点，强调产品可以进行机场客户流量的时事预测动态分析，并作出预警、提出解决方案，来吸引客户的目光。以产品的绝对优势来提升用户对于成本的投入。本产品购买方式便捷，操作方式简单，耗时短，能够做到人流量情况可视化。极高的性价比将帮助本团队产品快速赢得市场的认可，并且积累众多的产品拥护者。

（2） 促销策略

前期为了快速的打开市场，我团队将选取典型的试点，对本团队产品进行免费推广。通过向试点机场推荐安装本产品，机场对于产品的使用及使用情况反馈可以极大的增加本产品的社会公信力。其次是提供人性化的服务，免费为客户提供产品的操作培训以及产品情况的咨询，以便客户对于本产品能够有深入的了解，为每购买产品的客户建立完善的档案资料，以便进行产品使用情况的追踪与反馈，及时满足客户个性化的需求，提高客户忠诚度，吸引更多的人选择本产品，快速抢占市场。

6.4.3 优势分析

（1） 可行性分析

第一方面，“一对一”营销策略实现了企业与客户的互动，可以向客户提供一个简单可行的反馈系统，根据客户在使用本产品过程中所遇到的问题进行整理，快速制定出最合理的解决方案，满足客户的需求。

第二方面，“一对一”营销策略能够提供良好的反馈机制，可以随时根据客户的反馈和需求对产品和服务进行优化，且能够使客户与企业实现信息共享。

（2） 竞争优势

产品实施“一对一”营销策略在市场竞争中取得了极大的优势，它培养了“服务等于成功”这一经营理念，让客户满意成为了产品的最原始的动机。对客户而言，价值分为实体价值与虚拟价值两种，实体价值指的是真实存在的成本费用，虚拟价值指的是服务成本，这两种成本的市场竞争形成了强大的客户满意度倾向

值。对机场客流量大数据分析系统而言，做到服务领先、成本领先，可以在最大程度上提高竞争优势。本团队实施“一对一”营销策略在极大程度上满足了客户个性化的需求，能够使客户获得完美的使用体验，提升了客户对系统的信任度和忠诚度，在竞争中能够占据有利地位。

通过一对一的渠道与终端服务网络的建立，使得我们能够获得更多的客户的支持，我们也能够更多的了解客户的偏好和需求信息，但还有一个更大的与没有实施一对一的品牌的不同点，就在于我们将与客户的互动与对话转化成了可跟踪、能使用的信息。当我们把这些信息与企业的潜能相结合起来，就变成了企业的核心能力，这些信息就会在经过深层整合以后变成了一种知识，我们产品拥有了这种知识以后，我们就比我们的竞争队手更加了解我们的客户，特别对于客户群体划分出的个体的掌握，将会使我们做得更好，赢得客户更多更大的满意度。

6.5 服务营销策略

本产品属于互联网信息产业的高新技术产品，针对性、专业型强。针对客户需求，本产品有着严禁的“售前、售中、售后”服务保障体系。

6.5.1 售前服务

这一阶段，我们服务的主要目的是开展一系列刺激客户购买欲望的服务工作，协助客户做好规划和系统需求分析，使我们的系统最大限度的满足客户需要，同时也使客户的投资发挥出最大的综合经济效益。因此在这一阶段，我们主要有以下三大策略：一、提供情报，服务决策;二、解答疑问，引发需求；三、突出特点，稳定销售。 具体策略如下图：

图 6.3 售前策略

6.5.2 售中服务

在销售过程中，我团队根据客户要求，并对产品进行测试、调试，并对“机场客流量预测分析系统”操作人员进行培训，使相关人员掌握本产品的操作使用，保证系统能够发挥效能。同时为了激励和诱导客户购买本产品，在客户中树立良好的信誉和创造新的市场机会，我团队在售中还提供维修保养和产品管理相关知识的培训，以便客户提高系统的使用效率，一旦发生故障能够自行检修。

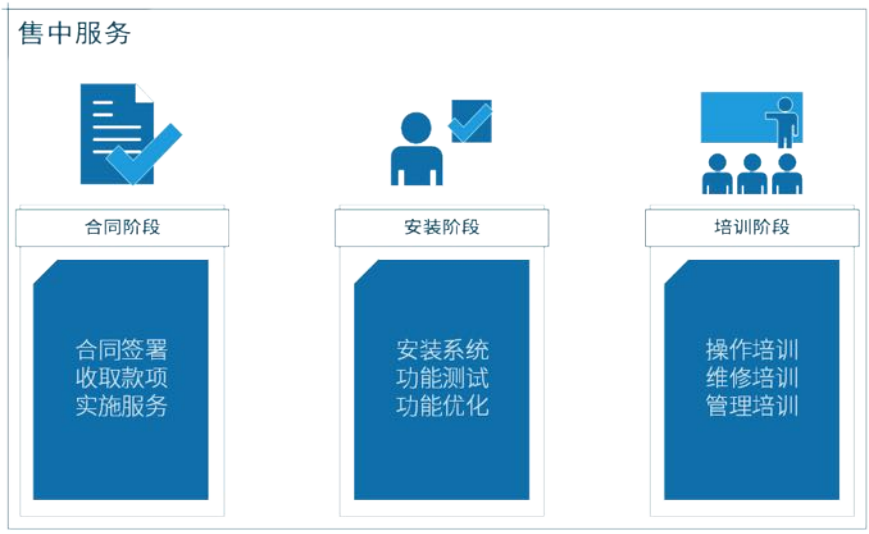


图 6.4 售中服务

6.5.3 售后服务

售后服务，主要包括对产品的定期维护、与客户进行有效沟通以及针对产品实际使用情况的优化阶段。

图 6.5 售后服务

机场客流量分析系统属于高新技术产品，我们的售后方式主要一下三种：技术支持、定期回访和免费的客服服务热线。

我们提供技术支持来帮助客户诊断并解决其在使用产品过程中出现的有明显症状的，可能由产品导致的技术问题，主要方式有：电话技术支持、上门技术支持等。

我们提供定期回访服务，通过对客户产品实际使用情况的了解，进行两方面的优化：一是技术上的维护，二是服务上的沟通。

我们开通 24 小时免费服务热线，客户在任何时候想要咨询或遇到相关产品技术问题，都会在短时间内得到我们的帮助和解决。

第7章 总结与展望

7.1 作品总结

身处在信息时代，对大数据的处理在当今各个行业中必不可少，除了分析统计之外，还可利用大数据进行预测，对于人流量大、安全要求高的场所实时预测更有必要，经过详细调查我们发现机场对此有很高的需求，但市场上几乎没有解决这些需求的产品，所以我们把目标定位为机场，用我们的技术为机场提供实时的客流量分析，以及预警和疏导方案的提供。

在挖掘经济获利点时，我们细致分析当前市场情况，除了为机场提供预测预警，我们还将人流信息推送给机场周边运力单位，提醒周围运力单位提供接送服务，加快机场人员疏导的同时，让旅客和运力单位获得多赢，除此之外，在后期数据量足够大时，我们会将服务推广给机场内部及周边商家，仿造以香港、新加坡为首的部分机场，逐步以机场服务为核心，将机场与商圈及周围设施融为一体，创建相应的商业模式。

在作品设计中，我们也遇到了许多难点，但通过努力，有找到了相应的解决方案，如下：

航班信息、安检口人流数据的部分异常值和缺失值对提取特征造成困难，必须寻找其他相关特征进行弥补。对此我们提出的解决方案：

（1） 以“登机口记录”为桥梁，将“Wi-Fi 点的旅客连接记录”和“航班记录”联系起来，提取出新特征；

（2） 删去异常值，用均值代替缺失值；

只用历史人流量进行建模预测，效果一般。解决方案：我们加入了航班表，选取有用的特征，提出无用的特征，再次建模，最后预测精度得到了显著的提高；

算法上将所有数据放在一个训练集中建模预测是精确度止步不前，即使添加多个新特征，也只有很小的提高。解决方案：我们发现不能提高的原因是，不同特征在相互干扰，故将整个训练集分模块训练，最后按权重进行融合。

7.2 展望

通过这段时间对作品的开发，我们也发现了作品的一些不足，对于预测精度，时间越长，预测的精度越低。为此，我们也做出了做一些展望：

（1） 扩展模型功能模块：开发并实现模型更多的功能，提高系统的数据处理、分析能力，展现更多的预测模块数据，从而进一步提升预测数据的质量，充分保证分析结果的真实性和准确性以及预测结果的可靠性，也给使用者提供更多的数据分析；

（2） 本作品以白云机场为落地案例，可在全国机场做推广和复制，为其他机场提供优化建议，使航站楼内的各类灯光电梯设施设备、值机柜台、商铺、广告位等安排更为合理，能更加精准、高效地调度这些资源和安排服务人员，减少机场该部分费用，增强机场安全；

（3） 推广到其他公共服务区：将数据搜集方式扩大化，兼容更多数据采集方式。将完善后的模型算法推广到类似地铁站、商圈等人流量大的区域，按照需求细化产品功能；

（4） 提高系统处理能力：从数据处理和预测数据两方面提高系统对于数据提纯的性能，支持处理更大量、更多维的数据；

（5） 改进和优化：作品功能还不够人性化和智能化，广泛听取更多使用者的意见，调整模块，增加人性化特色功能。降低数据挖据发开难度，降低维护成本。

（6） 提高系统存储能力：从数据处理效率和数据存储规模两方面提高系统存储性能，支持更大量级的数据存储能力。能够有效地保存之前几天已经发生的数据以及预测到的几天之内的数据，从而便于用户查询以及机场的管理。

相信通过我们的努力与思考，对产品做出进一步的研发与改进，使其能够为社会创造更多社会价值和财富价值，为人民带来更多便利，提高人们出行的舒适度和幸福感。

参 考 文 献

[1]Jiawei Han, Micheline Kamber and Jian Pei. Data Mining Concepts and Techniques, Third Edition. Beijing: China Machine Press, 2012: 288-291 (韩家炜, 坎伯, 裴建. 数据挖掘 : 概念与技术, 第3版= Data Mining: Concepts and Techniques, Third Edition : 英文[M]. 北京: 机械工业出版社, 2012: 288-291)

[2] Liu Zhihui, Zhang Quanling. Journal of Zhejiang University. Research overview of big data technology. 10.3785/j.issn.1008-973X. 2014.06.01

[3]Christian Tominski, Event-Based Concepts for User-Driven Visualization [J] Information Visualization, 2011,10(1):65-81

[4] Keim DA, Kriegel HP. Visualization techniques for mining large databases:A coparison [J] Trans. on knowledge and Data Engerneering 1996,8(6)923-938

[5]Hey T, Gannon D, Pinkelma J. The future of data-intensive science [J]. Computer.2012,45(5):81-82

[6]Kriegel H P, Pfeifle M. Hierarchical Density-Based Clustering of Uncertain Data[C]// Eleventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Chicago, Illinois: ACM, August. 2005: 689-692

[7]Viswanath P, Pinkesh R. l-DBSCAN: A Fast Hybrid Density Based Clustering Method[C]// International Conference on Pattern Recognition. HongKong: IEEE, 2006: 912-915

[8]Hou J, Gao H, Li X. DSets-DBSCAN: A Parameter-Free Clustering Algorithm[J]. IEEE Transactions on Image Processing, 2016, 25(7): 3182-3193

[9]Amini A, Saboohi H, Wah T Y. A Multi Density-Based Clustering Algorithm for Data Stream with Noise[C]// IEEE, International Conference on Data Mining Workshops. Dallas, Texas: IEEE, 2013: 1105-1112

[10]Birant D, Kut A. ST-DBSCAN: An algorithm for clustering spatial-temporal data[J]. Data & Knowledge Engineering, 2007, 60(1): 208-221

[11]Ankerst M, Breunig M M, Kriegel H P, et al. OPTICS: ordering points to identify the clustering structure[J]. Acm Sigmod Record, 1999, 28(2): 49-60

[12]Patwary M A, Palsetia D, Agrawal A, et al. Scalable parallel OPTICS data clustering using graph algorithmic techniques[C]//International Conference for High Performance Computing, Networking, Storage and Analysis. New York: ACM, 2013

[13]Goyal P, Kumari S, Kumar D, et al. Parallelizing OPTICS for multicore systems[C]// ACM COMPUTE. New York: ACM, 2014: 1-6

[14]Deng Z, Hu Y, Zhu M, et al. A scalable and fast OPTICS for clustering trajectory big data[J]. Cluster Computing, 2015, 18(2): 549-562

[15]Špitalský V, Grendár M. OPTICS-Based Clustering of Emails Represented by Quantitative Profiles[M]// Distributed Computing and Artificial Intelligence. Berlin: Springer, 2013: 53-60

[16]Kalita H K, Bhattacharya D K, Kar A. A New Algorithm for Ordering of Points to Identify Clustering Structure Based on Perimeter of Triangle: OPTICS(BOPT)[C]// International Conference on Advanced Computing and Communications. Guwahati, Assam: IEEE Computer Society, 2007: 523-528

[17]Ankerst M. OPTICS: ordering points to identify the clustering structure[J]. Acm Sigmod Record, 1999, 28(2): 49-60

[18]Zaharia M, Chowdhury M, Franklin M J, et al. Spark: cluster computing with working sets[C]// Usenix Conference on Hot Topics in Cloud Computing. Boston :USENIX Association, 2010: 1765-1773

[19]Bharill N, Tiwari A, Malviya A. Fuzzy Based Scalable Clustering Algorithms for Handling Big Data using Apache Spark[J]. IEEE Transactions on Big Datam, 2016, 4(2): 339-352

[20]Li J, Li D, Zhang Y. Efficient Distributed Data Clustering on Spark[C]// IEEE International Conference on CLUSTER Computing. Chicago: IEEE Computer Society, 2015: 504-505
[21]Sinha A, Jana P K. A novel K-means based clustering algorithm for big data[C]// International Conference on Advances in Computing, Communications and Informatics. Jaipur: IEEE, 2016: 1875-1879
[22]Jin C, Liu R, Hendrix W, et al. A Scalable Hierarchical Clustering Algorithm Using Spark[C]// IEEE First International Conference on Big Data Computing Service and Applications. San Francisco Bay: IEEE, 2015: 418-426
[23]Sarazin T, Azzag H, Lebbah M. SOM Clustering Using Spark-MapReduce[C]// IEEE International Parallel & Distributed Processing Symposium Workshops. Phoenix: IEEE Computer Society, 2014: 1727-1734
[24]Conrad J G, Al-Kofahi K, Zhao Y, et al. Effective document clustering for large heterogeneous law firm collections[C]// Proceedings of the 10th international conference on Artificial intelligence and law. Bologna: ACM, 2005: 177-187
[25]UCI Machine Learning Repository [DB/OL].
http://archive.ics.uci.edu/ml
[26]Wikipedia Weka (machine learning) [CP/OL].
http://en.wikipedia.org/wiki/Weka,2010
[27]xj1986. MR-DBSCAN [EB/OL]. [2013-5-15].
https://github.com/xj1986/MR-DBSCAN

附 件

1 相关专利

1.1 基于 Spark 内存计算大数据平台的 OPTICS 点排序聚类方法

AJ165447_5447_XSQ_20161233

中华人民共和国国家知识产权局

401121

重庆市渝北区红锦大道 498 号佳乐紫光大厦 7 楼
重庆市恒信知识产权代理有限公司 刘小红,高敏

发文日：
2016 年 12 月 08 日

申请号或专利号：201611120326.3　　发文序号：2016120800677930

专 利 申 请 受 理 通 知 书

根据专利法第 28 条及其实施细则第 38 条、第 39 条的规定，申请人提出的专利申请已由国家知识产权局受理。现将确定的申请号、申请日、申请人和发明创造名称通知如下：

申请号：201611120326.3
申请日：2016 年 12 月 08 日
申请人：重庆[redacted]大学
发明创造名称：基于 Spark 内存计算大数据平台的 OPTICS 点排序聚类方法

经核实，国家知识产权局确认收到文件如下：
专利代理委托书 每份页数:2 页 文件份数:1 份
说明书附图 每份页数:5 页 文件份数:1 份
说明书 每份页数:17 页 文件份数:1 份
实质审查请求书 每份页数:1 页 文件份数:1 份
发明专利请求书 每份页数:4 页 文件份数:1 份
权利要求书 每份页数:3 页 文件份数:1 份 权利要求项数： 8 项
说明书摘要 每份页数:1 页 文件份数:1 份

提示：
1. 申请人收到专利申请受理通知书之后，认为其记载的内容与申请人所提交的相应内容不一致时，可以向国家知识产权局请求更正。
2. 申请人收到专利申请受理通知书之后，再向国家知识产权局办理各种手续时，均应当准确、清晰地写明申请号。
3. 国家知识产权局收到向外国申请专利保密审查请求书后，依据专利法实施细则第 9 条予以审查。

审 查 员：自动受理　　审查部门：专利局初审及流程管理部

中华人民共和国国家知识产权局 专利申请受理章 (19)

200101 2010.4　纸件申请，回函请寄：100088 北京市海淀区蓟门桥西土城路 6 号 国家知识产权局受理处收
电子申请，应当通过电子专利申请系统以电子文件形式提交相关文件。除另有规定外，以纸件等其他形式提交的文件视为未提交。

1 / 1

1.2 基于 Spark 大数据平台的三支决策不平衡数据过采样方法

AJ165496_5496_XSQ_20161256

中华人民共和国国家知识产权局

401121

重庆市渝北区红锦大道 498 号佳乐紫光大厦 7 楼
重庆市恒信知识产权代理有限公司 刘小红

发文日：
2016 年 12 月 29 日

申请号或专利号：201611244051.4　　发文序号：2016122901384850

专利申请受理通知书

根据专利法第 28 条及其实施细则第 38 条、第 39 条的规定，申请人提出的专利申请已由国家知识产权局受理。现将确定的申请号、申请日、申请人和发明创造名称通知如下：

申请号：201611244051.4
申请日：2016 年 12 月 29 日
申请人：重庆[redacted]大学
发明创造名称：基于 Spark 人数据平台的三支决策不平衡数据过采样方法

经核实，国家知识产权局确认收到文件如下：
说明书附图 每份页数:3 页 文件份数:1 份
专利代理委托书 每份页数:2 页 文件份数:1 份
权利要求书 每份页数:4 页 文件份数:1 份 权利要求项数： 9 项
实质审查请求书 每份页数:1 页 文件份数:1 份
说明书摘要 每份页数:1 页 文件份数:1 份
说明书 每份页数:14 页 文件份数:1 份
发明专利请求书 每份页数:5 页 文件份数:1 份

提示：
1. 申请人收到专利申请受理通知书之后，认为其记载的内容与申请人所提交的相应内容不一致时，可以向国家知识产权局请求更正。
2. 申请人收到专利申请受理通知书之后，再向国家知识产权局办理各种手续时，均应当准确、清晰地写明申请号。
3. 国家知识产权局收到向外国申请专利保密审查请求书后，依据专利法实施细则第 9 条予以审查。

审 查 员：自动受理　　审查部门：专利局初审及流程管理部

专利申请受理章

200101
2010.4

纸件申请，回函请寄：100088 北京市海淀区蓟门桥西土城路 6 号 国家知识产权局受理处收
电子申请，应当通过电子专利申请系统以电子文件形式提交相关文件。除另有规定外，以纸件等其他形式提交的文件视为未提交。

1 / 1