



电信运营商大数据基础能力建设与实践

汤劲松, 蔡韵, 王晓亮, 傅一平, 汤潇巍, 陈永刚, 张挺
(中国移动通信集团浙江有限公司, 浙江 杭州 310016)

摘要: 以中国移动通信集团浙江有限公司(以下简称浙江移动)为例, 从平台、数据和技术 3 方面阐述了运营商的大数据能力, 并结合企业实践和探索, 分享了浙江移动在移动信用评分和智能选址产品的案例。

关键词: 大数据; 信用评分; 选址

中图分类号: TP311

文献标识码: A

doi: 10.11959/j.issn.1000-0801.2018037

Construction and practice of big data fundamental ability for telecom operators

TANG Jinsong, CAI Yun, WANG Xiaoliang, FU Yiping,
TANG Xiaowei, CHEN Yonggang, ZHANG Ting
China Mobile Group Zhejiang Co., Ltd., Hangzhou 310016, China

Abstract: Take Zhejiang Mobile as an example, the big data capability was demonstrated from three aspects: platform, data and technology. Combined with the enterprise practice and exploration, industry cases were shared in terms of credit score and site selection.

Key words: big data, credit score, site selection

1 引言

随着互联网业务和应用的迅猛发展以及移动互联网的爆炸式增长, 电信运营商客户基础属性、行为数据、信令数据和终端数据等海量数据的存储与分析日益成为电信运营商的重要挑战。大数据技术的出现与机器学习的发展为电信运营商深挖数据提供了新的技术手段, 同时也为其更好地服务客户提供了新的机遇^[1]。

运营商通过大数据平台的建设、海量数据的

加载、丰富的画像标签沉淀以及上层机器学习算法的深入研究, 针对几大相关行业进行了深入探索和实践^[2], 也推出了系列创新产品, 在支撑金融风险控制、用户征信领域和商业决策等方面发挥着越来越重要的作用。

2 大数据基础能力

2.1 平台架构

中国移动通信集团浙江有限公司(以下简称浙江移动)大数据平台经过两期建设, 已建成

Hadoop、MPP、流处理和内存数据库等平台，集群规模已达1 400+，日均处理离线数据100多TB、实时数据千亿条，融合了OSS（operation support system，运营支撑系统）、BOSS（business support system，业务支撑系统）、MSS（management support system，管理支撑系统）三域内部数据+互联网外采多维数据^[3]，为大数据应用提供有力的PaaS能力支撑。浙江移动大数据平台如图1所示。

2.2 数据标签能力

运营商在数据方面具有天然优势，拥有垄断的通话关系网数据，最实时、最全面的位置数据，最丰富的上网行为数据，最全面的终端行为数据，最权威的用户身份信息数据，最详细的通信消费数据^[4]。以下是各类数据的特点和价值。

(1) 通话关系网数据

- 通话社交关系数据：基于通话交往圈的大小、主被叫及时间规律，就掌握了任何一个用户的社交特征，比如某人的影响力，人与人之间的亲密程度，人群之间的上下属关系，甚至可以得到人脉路径。

- 通话时序数据：通话的频次、时序、时长及对端等重要数据，又可以用来挖掘用户的性格甚至身份特质、作息规律等。

(2) 位置数据

- 全方位的信令位置数据：运营商的位置数据不单单是需要通话、上网、短信才能获取，而是通过定期基站切换、周期性位置上报而直接获取用户全方位位置数据，不依赖用户是否登录使用某一手机应用，这是较GPRS（general packet radio service，通用分组无线服务）技术定位最大的优势。
- 实时性强：具备实时的流处理分钟级的用户位置数据，可用于各类实时位置服务应用。

(3) 上网数据

- 拥有移动用户手机上网的所有行为数据，经过DPI（deep packet inspection，深度分组检测）技术内容解析，互联网商品最深可到7级，App（application，应用程序）识别9 000个。
- 用户搜索行为数据：用户手机上的搜索行

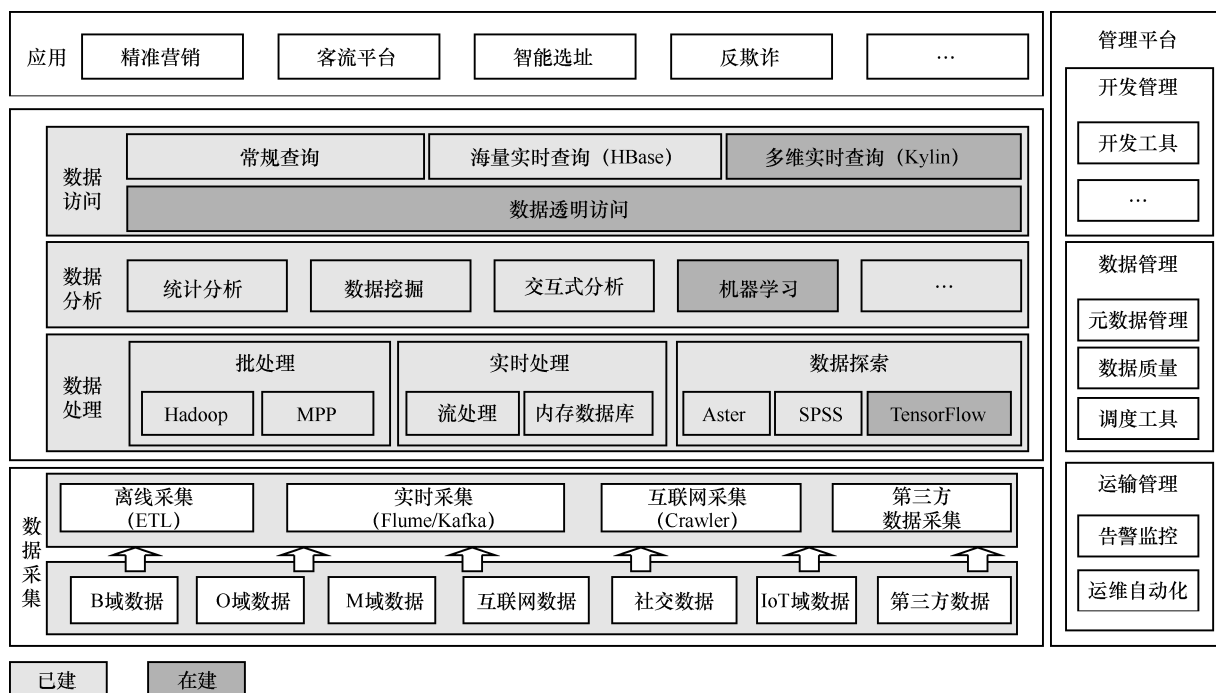


图1 浙江移动大数据平台架构



为是用户近期最强的需求指向，只要进行语义解析并结合一定的知识库，就可以在精确营销中发挥巨大价值。

(4) 用户终端信息

- 用户使用的终端信息：包括型号、品牌、换机频次、品牌忠诚度等，可以掌握终端市场，可以从终端看用户档次。
- 用户换机轨迹：可以用于分析终端品牌之间的流向。

(5) 身份信息数据

- 实名制普遍推广后，运营商的身份数据质量极高，包括身份证号码、姓名等真实信息。

(6) 通信消费数据

- 用户每月的通信支出、支出分布、充值情况、充值频次等，一定程度上可以体现用户的消费习惯、收入水平。

基于用户画像维度和运营商数据特点，标签体系从上往下结构化分为七大类 60 小类，分别从客户的基础属性、通信属性、习惯偏好、社交关系、App 行为、线上事件和位置轨迹进行展现，标签规模突破了 10 万个。浙江移动标签体系如图 2 所示。

2.3 技术能力

(1) 云爬虫平台

为丰富数据类型和规模，浙江移动自主研发

了云爬虫平台来获取外部数据并提供分词和自然语言的解析能力。云爬虫平台可以抓取一切调用 API 访问的网站内容，支持定制化的垂直爬取需求，租户能够自主指定各主流网站所要爬取的数据。目前，云爬虫平台可实现日均 1 亿 URL 的采集量，有效爬取数据存储能力大于 500 TB，日均处理请求超过 1 500 万条，日均爬取数据量超过 2 TB。目前已采集了工商企信、个人征信和 POI (point of interest, 兴趣点) 等外部信息。

(2) MR 精准定位

MR (measurement report, 测量报告) 精准定位模型利用 MR 和 OTT (over the top, 通过互联网向用户提供各种应用服务) 数据，通过定位算法对移动终端用户进行精准定位，能够精确到 50 m 以内，可精准输出目标用户的位置规律，为解读用户的位置行为研究提供数据保障。当前，浙江移动的 MR 精准定位模型已经开展全量数据处理和试用，定位精度从基站的 300~500 m 可有效提升到 50~100 m，定位误差达到商用要求。

(3) DPI 深度解析技术

浙江移动运用 DPI 针对每天 500 亿网页进行解析，已解析 9 000 多款 App，涵盖金融、购物、游戏、娱乐、阅读等 19 大类，网页内容解析后，识别图书、视频、音乐、资讯、商品、点评等 26 大类 8 000 细类。

二级体系	基础属性	通信属性	习惯偏好	社交关系	位置轨迹
	• 社会属性 • 家庭属性 • 工作属性	• 渠道接触 • 通信使用 • 通信消费 • 业务订购 • 终端使用 • 通信事件 • 通信特征	• 生活习惯 • 饮食习惯 • 消费习惯 • 交通习惯 • 娱乐习惯	• 生活交往圈 • 工作交往圈 • 社交属性	• 休闲娱乐 • 生活服务 • 交通设施 • 节假日出行 • 展会 • 其他
	App行为			线上行为	
	• 办公商务 • 便捷生活 • 儿童亲子 • 购物优惠 • 婚恋交友 • 交通导航 • 教育学习	• 金融理财 • 旅游酒店 • 其他杂项 • 摄影摄像 • 实用工具 • 通信社交 • 系统安全	• 新闻阅读 • 影音视听 • 游戏娱乐 • 运动健康 • 主题壁纸	• 本地生活 • 通信营业厅 • 论坛帖子 • 旅游 • 美食 • 票务 • 商品	• 视频 • 图片 • 图书 • 微博 • 音乐 • 应用 • 邮件 • 游戏 • 资讯 • 其他

图 2 浙江移动标签体系

3 运营商大数据案例分享

3.1 移动信用评分应用

3.1.1 项目背景

随着 P2P 市场的野蛮生长和消费金融在中国的兴起,个人征信市场的需求大幅爆发。在 2015 年 1 月,中国人民银行(以下简称央行)发布通知要求芝麻信用、腾讯征信等 8 家公司机构做好个人征信的准备工作,这一举动说明征信业正式向市场化发展,同时,也标志着中国征信业正式进入 2.0 时代。2.0 时代的个人征信业存在两个特征:一是征信机构的多元主体化,征信牌照的发放使更多的机构进入市场,这有利于丰富和完善个人征信的服务体系;二是技术产品的创新,随着大数据的到来,数据规模越来越多,在征信数据源和征信模型方面都有较多的创新和突破,征信体系有很多的改善空间。

中国移动拥有全国近 10 亿用户的通信行为数据、用户位置数据及全网移动互联网数据以及公共事业单位拥有的日常支付消费数据。这些数据拥有全面性、及时性、形式多样性和可信度高等优势。中国移动可以依据多年积累的大数据技术,充分应用这些数据搭建个人征信评分体系,用于服务自身传统业务运营以及提供给其他征信金融机构作为个人信用评估标准或征信数据源的补充。

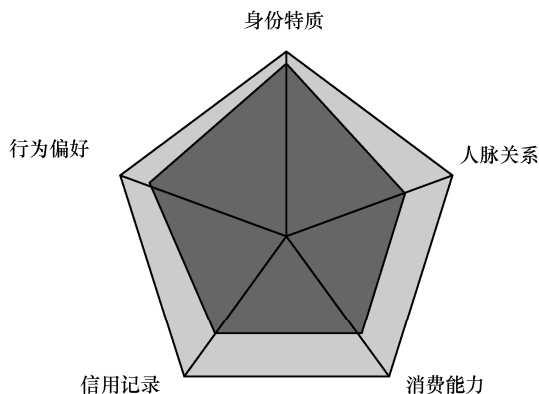


图3 移动信用评分体系

3.1.2 模型介绍

浙江移动在充分借鉴 FICO、芝麻信用等行业标杆先进做法的基础上,基于浙江移动的大数据平台,充分整合客户身份、通话、上网和位置等核心数据,利用决策树、逻辑回归和随机森林等数据挖掘模型,从身份特征、消费能力、行为偏好、人脉关系和信用记录 5 个维度构建客户信用评分模型,实现了对客户信用的客观评价,并通过欠费、坏账、行业违约等数据的验证,不断迭代和优化模型。移动信用评分体系如图 3 所示,移动信用评分建模过程如图 4 所示。

3.1.3 模型效果

模型在测试样本下预测的准确率如图 5 所示,其中 TPR 为真阳率, FPR 为假阳率。从图 5 可以看出,该模型 ROC 曲线的 AUC 值能达到 80%,说明模型预测效果不错。模型在测试样本下 K-S 曲线如图 6 所示。从测试样本的 K-S 曲线来看,个人信用分的最大 K-S 值能达到 0.7,区分度较好。

3.1.4 产品应用

浙江移动首选欠费免停机场景作为试点业务启动,针对不同的星级客户给予了不同的信用额度(透支额度)。当客户产生的欠费在信用额度内时,对用户进行停机操作。通过该试点业务,在当前的星级体系下找出星级较低但是信用分较高的群体,赋予他们更高的信用额度,

身份特质	• 根据用户的注册信息、家庭信息等进行评分
行为偏好	• 根据用户语言、流量、省际漫游、出国漫游国际长途等信息进行评分
信用历史	• 根据用户历史停机、欠费、坏账信息及在合作方、公检法等机构的不良记录进行评分
人脉关系	• 根据用户交往圈大小、在交往圈中的影响力、交往圈整体信用情况进行评分
消费能力	• 根据用户语音、流量、积分等的使用情况,结合各类App的使用进行评分

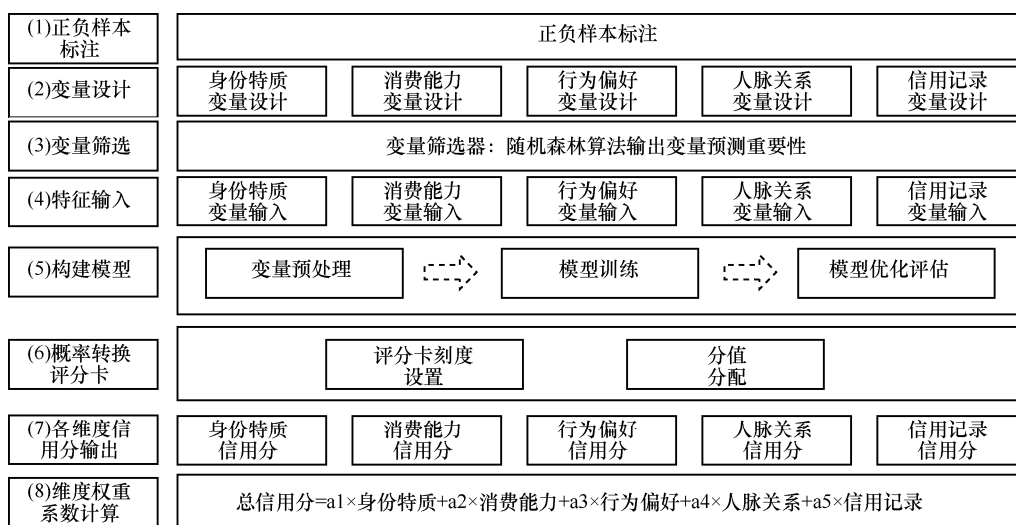


图4 移动信用评分建模过程

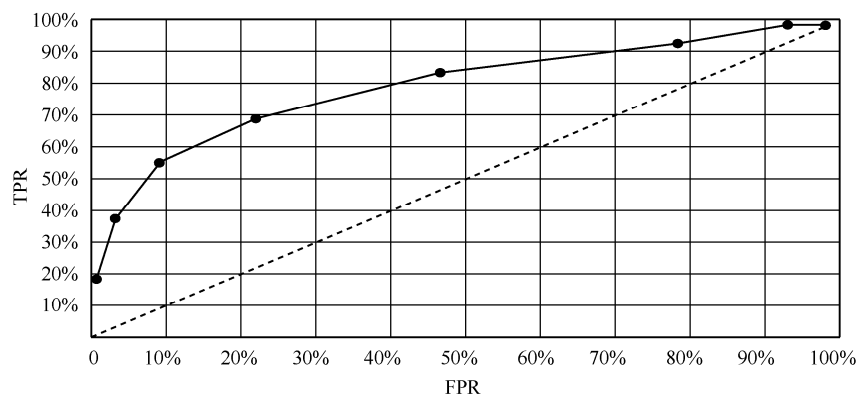


图5 模型在测试样本下预测的准确率

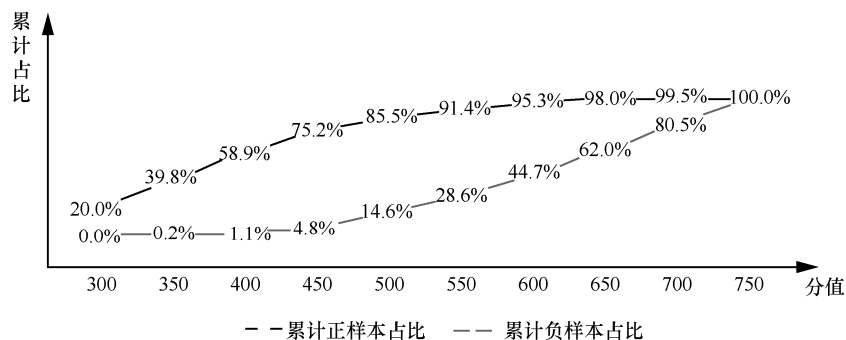


图6 模型在测试样本下 K-S 曲线

以减少用户的停机可能，提升收入同时控制欠费风险。

(1) 目标客户捞取

从客户星级和信用分两个维度综合考虑欠费率情况，最终捞取信用分在 700 分以上，但星级

在五星以下的客户，为其提供免停机服务。

(2) 信用额度调整

针对目标客户群体，根据不同的信用分等级，将其提升至不同档次的信用额度。高信用低星级用户信用额度调整规则如图 7 所示。

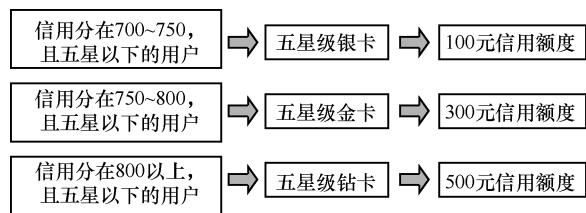


图7 高信用低星级用户信用额度调整规则

(3) 目标用户体验

基于信用的话费透支额度提升体验活动于8月1日开始在浙江多地开展，在目标用户中随机选取用户进行外呼，8月份累计外呼25.6万人，其中，接通11.3万人，成功7.1万人，最终转化率为27.6%。在8月份期间，共有2.2万人使用了话费透支服务（即8月份发生过欠费），占有参与活动客户的33%。

应用成效显著，主要体现在以下方面。

(1) 停机得到明显改善

由于信用额度的提升，停机用户数下降98%以上，客户的服务感知得到了提升和改善。

(2) 收入提升显著

为了排除自然增长等因素的影响，与全省同条件且未参加活动的用户相比，参加活动的用户8月份ARPU（average revenue per user，用户平均收入）提升2.17元。

(3) 欠费风险较低

活动客户8月的欠费率为0.49%（39天后），显著小于全网客户同期的欠费率1.1%。

3.1.5 产品展望

基于信用评分应用浙江公司当前已规划了六大基于信用分的场景，对内逐步从欠费免停机、国内漫游免预存、信用购机方面开展；对外将从消费贷、免押金骑行和酒店信用住等领域拓展合作。

3.2 智能选址

3.2.1 研究现状

选址一直是实体商业、城市规划、基础设施建设领域的重要课题。目前，在学术界一直有非

常广泛的研究和探索。才鑫等人^[5]主要基于移动通信用户的上网数据、通话数据及位置更新数据对用户社会化行为进行有效分析，通过对具有相似兴趣的用户轨迹进行聚类、预测及可视化显示，进行商业选址，而没有将交通、环境、竞争等因素考虑进去；同时，并不能解决系统自动决策网点位置的问题。黎夏等人^[6]将遗传算法、蚁群算法、粒子群算法等人工智能算法应用到多点选址问题中；黎海波等人^[7]利用多目标粒子群优化算法和区域形状变异算法相结合来解决复杂的空间选址问题，取得了较好的效果。但这些都是基于GIS（geographic information system，地理信息系统）体系的POI（point of information，信息点）数据体系来构建的选址体系，都没有考虑人类活动的因素，造成选址结果与人们真实的需要往往存在一定差距。随着信息技术的发展，手机信令数据、社交网络数据等大数据渐渐发挥出重要的作用，大数据技术就是将原本看起来很难利用的数据抽丝剥茧，发现其中的真正内涵^[8]，整合多方数据进行选址也成为可能。

浙江移动大数据体系目前已经整合用户特征数据、用户通信行为数据、用户位置数据、用户互联网行为数据以及外部的POI数据和商业网点数据，形成了涵盖地理空间、产业和居民的全量数据体系，具备了进行智能选址的数据基础。同时，鉴于目前大量的选址研究还仅仅是停留在GIS热力图展示的定性化研究的现状，本文将基于浙江移动大数据建模和分析平台能力，结合全量数据体系，提出基于地理数据、产业数据和用户数据的智能选址体系，对传统的单纯基于POI或“人”的选址方法进行优化补充，构建智能推荐和决策的选址系统。

3.2.2 选址方案

整合信令位置、客户特征、POI 3方面数据，采用分布式+MPP并行数据处理架构，构建基于地图网格分析模型，实现智能选址应用。对外

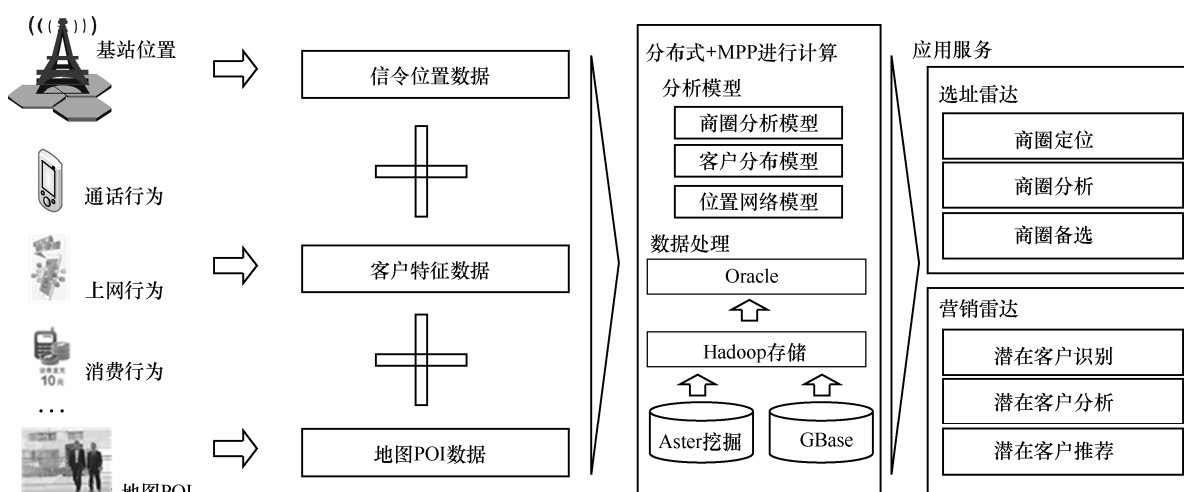


图8 智能选址产品技术

为规模性的连锁机构或高价值、低频次的大型商场提供商圈初筛、商圈分析、商圈对比等功能，辅助商家决策更加合适的选址点。对内提供区域客户对实体渠道、摆摊方式的需求评估、周边客户的效能评估、竞争环境监测与分析等能力，辅助评估备选点是否适合开实体渠道或进行摆摊。对内对外均可通过特征选取、导入客群、相似客群挖掘 3 种方式提取潜在客户，采用图形化界面展示目标客户的特征和分布，便于选择最佳营销触点方式和营销策略，力求更科学的精准营销，智能选址产品技术如图 8 所示。

3.2.3 应用效果

选址应用在对外应用方面，已经与房地产、商场零售等各行业商家开展选址、拓客合作。在对内应用方面，已开始支持中国移动的渠道布局决策、加盟厅评估和审核等工作。以加盟厅评估为例，原来加盟厅审批时，地市移动需人工采集和上报厅店周边人口、竞争、商圈业态等信息，现在由应用直接提供，简化了流程，开店时间缩短至原来的 1/3，低效厅控制在 10% 以下。在对外应用上，通过对点评评分排名前 10 的快餐店进行位置与模型智能推荐的排名前 10 位置进行对比，有 7 个点模型推荐的位置与实际的位置基本一致，同时实际排名靠前的另外 3 个点也都在模型推荐

的理想开店地址。因此，能够明显地看到通过多方数据的结合以及智能推荐模型的构建，能够帮助餐饮行业实现智能化一键选址。

4 结束语

运营商在数据方面天生拥有得天独厚的优势，未来将在人工智能、行业理解和产品创新上不断探索前进、努力转型，使运营商的数据和能力在越来越多的行业中发挥举足轻重的作用^[9]。

参考文献：

- [1] 吴军. 大数据和机器智能对未来社会的影响[J]. 电信科学, 2015, 31(2): 7-16.
WU J. Big data, machine intelligence and their impacts to the future world [J]. Telecommunications Science, 2015, 31(2): 7-16.
- [2] 刘春, 邹海峰, 向勇. 大数据环境下电信数据服务能力开放研究[J]. 电信科学, 2014, 30(3): 156-161.
LIU C, ZOU H F, XIANG Y, et al. Research on telecom data service open ability under the environment of big data [J]. Telecommunications Science, 2014, 30(3): 156-161.
- [3] 于鹏. 数据仓库与大数据融合的探讨[J]. 电信科学, 2015, 31(3): 166-170.
YU J. Discussion on integration of data warehouse and big data [J]. Telecommunications Science, 2015, 31(3): 166-170.
- [4] 傅一平, PK BAT. 运营商大数据其实更有价值[J]. 广告主, 2016(4).
FU Y P, PK BAT. Big data operators is more valuable in fact [J]. Advertisers, 2016(4).
- [5] 才鑫, 姜国强. 基于移动大数据的商业选址系统的研究与实

现[J]. 移动通信, 2015(13): 1.

CAI X, JIANG G Q. Research and implementation of commercial location system based on mobile big data[J]. Mobile Communication, 2015(13): 1.

- [6] 黎夏, 叶嘉安. 遗传算法和 GIS 结合进行空间优化决策[J]. 地理学报, 2004, 59(5): 745-753.

LI X, YE J A. Optimal spatial search using genetic algorithms and GIS[J]. Acta Geographica Sinica, 2004, 59(5): 745-753.

- [7] 黎海波, 黎夏, 刘小平, 等. 多目标粒子群算法与选址中的形状优化[J]. 遥感学报, 2008, 12(5): 724-733.

LI H B, LI X, LIU X P, et al. Particle-swarm optimization for site selection with contiguity constraints[J]. Journal of Remote Sensing, 2008, 12(5): 724-733.

- [8] 吴志峰, 柴彦威, 党安荣, 等. 地理学碰上“大数据”: 热反应与冷思考[J]. 地理研究, 2015, 34(12): 2207-2221.

WU ZH F, CHAI Y W, DANG A R, et al. Geography interact with big data: Dialogue and reflection[J]. Geographical Research, 2015, 34(12): 2207-2221.

- [9] 廖建新. 大数据技术的应用现状与展望[J]. 电信科学, 2015, 31(7): 7-18.

LIAO J X. Big data technology: current applications and prospects [J]. Telecommunications Science, 2015, 31(7): 7-18.

[作者简介]



汤劲松 (1968-), 男, 博士, 中国移动通信集团浙江有限公司高级工程师、大数据中心经理, 主要研究方向为大数据模型、大数据行业研究、运营商大数据挖掘及商业变现等。



蔡韵 (1980-), 女, 中国移动通信集团浙江有限公司工程师、数据建模师, 主要研究方向为数据建模和数据管理。



王晓亮 (1984-), 男, 中国移动通信集团浙江有限公司大数据中心工程师、产品经理, 主要研究方向为用户标签和行为。

傅一平 (1976-), 男, 博士, 中国移动通信集团浙江有限公司高级工程师、大数据中心数据管理部经理, 主要从事大数据系统、大数据建模、大数据营销、大数据行业研究、运营商大数据挖掘及商业变现等方面的研究工作。

汤潇巍 (1990-), 男, 中国移动通信集团浙江有限公司工程师、数据建模师, 主要研究方向为数据建模。

陈永刚 (1981-), 男, 中国移动通信集团浙江有限公司工程师、产品经理, 主要研究方向为大数据产品及模型。

张挺 (1987-), 男, 中国移动通信集团浙江有限公司工程师、产品经理, 主要研究方向为大数据产品。