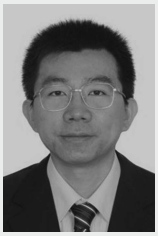
中国电信大数据应用实践

Application of big data in China Telecom



张宇中(1969-),男,中国电信股份有限公司云计算分公司首席数据分析师、大数据分析顾问,主要研究方向为消费者研究、互联网网民行为分析和数据挖掘、新媒体传播及媒介价值研究、网络营销效果评估优化、汽车数字营销。

李名洋(1983-),男,中国电信股份有限公司云计算分公司数据分析师,主要负责大数据分析、模型搭建应用、行业大数据研究等工作。

中图分类号:TP399 文献标识码:A doi: 10.11959/j.issn.2096-0271.2016036

* 本文为2015中国大数据技术大会(BDTC)演讲约稿

1 引言

大数据的应用随着计算技术的进步、互联网的爆发、科学计算的需求而高速发展。各类互联网巨头公司积累了大量运营、用户和交易数据，并筹建了大量的运算资源。它们的各类商业目的推动了大数据处理技术的发展。

对中国电信运营商而言，三网总的活跃移动连接超过10亿，其中超六成终端为智能终端，每天各类应用和业务被使用，持续产生大量的数据流。用户通过智能终端的通信和数据业务使用各类应用，使移动网络成为大数据存储、流动的天然载体。运营商不仅拥有传统的用户基础信息、网络数据，还有通过管道功能获取的用户互联网活动数据，用户信息全面真实。

而这些数据的利用面临着诸多的问题。从数据的归属权和隐私控制方面看，数据拥有权和使用权的划分、用户授权方式、法律风险的防控等，对大数据行业的发展方向有较大的影响；从数据的有效性看，在大量数据中寻找关联信息并验证其有效性，是非常巨大的工作量；从业务逻辑看，对于运营商来讲，将原本用于经营的数据产生机制用于大数据领域，需要从硬件、软件、人才诸方面进行调整，甚至可能面临大的变革。

大数据的应用问题不仅仅是一个企业内部的事情，也是整个行业乃至跨行业的事情。从软硬件方面看，大数据应用涉及硬件设施、基础软件、应用软件和信息服务等方面；从数据生产流程看，大数据应用涉及数据生成与采集、数据存储、数据处理和数据应用。在运营商核心数据资源的外围，聚集着大量各类型、各行业的支撑公司、上下游企业和合作伙伴。

2 电信大数据发展概述

2.1 中国电信的大数据业务

中国电信大数据业务的开展依托于中国电信云计算分公司，由集团市场部直接管理。中国电信2014年开始启动全国大数据的集约化运营，着力推进全集团数据汇集和发掘应用，实现大数据应用产业化、规模化发展。根据集团规划，云计算分公司作为大数据运营支撑单位，承担大数据产品集约开发、运营、销售、服务和经营支撑工作。

2.2 云计算分公司大数据工作的主要内容

中国电信大数据数量巨大、来源分散、格式多样，对系统的数据处理能力和分析挖掘能力提出了巨大挑战，需要新技术将庞杂无序的数据进行清洗、处理、分析和集成，变成有用的信息，与行业应用融合产生价值。目前这主要涉及以下4个方面的工作内容。

- 建设大数据能力平台。实现全网数据集约（接入、计算、存储）及数据资产集中管理维护。
- 建设海量数据处理所需的五大基础能力。包括数据传导（被动/主动接入能力、数据传导、数据开放）、存储（结构化、非结构）、计算（批量、流式）、安全运营（提供对数据、系统安全运营保障的手段）、资源调度（协同协调、资源隔离、能力配额）。
- 数据生产线技术架构设计。适应电信大数据两种业务数据模型，包括批量—调度系统：基础表、母表、子表、基础服务层；实时—消息系统：基础拓扑、融合拓扑、基础服务层。

● 产品应用体系设计。选择市场需求明确、市场规模大、应用模式清晰、适合电信大数据特点的领域建设产品应用平台。

中国电信大数据能力产品与应用体系如**图1**所示。

3 电信大数据的特点与开发利用

3.1 电信大数据的构成与特点

(1)中国电信大数据的构成

中国电信的数据优势在于数据的广度和深度。中国电信具有海量数据基数，包括2亿手机用户和1.5亿宽带用户（覆盖了全国70%的宽带用户上网份额）产生的数据。此外，还有IPTV、Wi-Fi热点数据。这些数据涵盖运营商全业务形态。同时，中国电信还拥有大量第三方基础合作数据。中国电信自有数据主要包括IT类数据、网络类数据、信令数据和终端数据。合作数据包括地图POI（兴趣点）数据、金融征信类数据、行业数据等。

(2)电信大数据的特点

中国电信拥有大量真实的用户。真实有效的数据能够支撑可信度高的分析与结论，还可进行多维度精细用户群体分析。中国电信作为互联网接入服务提供商，承载了国内电信用户各类业务数据，涵盖通信数据、业务数据、互联网数据、信令数据等方面，提供全方位的服务，数据应用的可靠性高。同时，中国电信用户本身样本的覆盖具有无偏差的特点，可以有效保证分析的准确度。

3.2 电信大数据的开发原则

完善的隐私保护、提供安全可靠的服务、平台级的开发能力、支撑行业企业发展、构建健康大数据生态环境，是电信大数据开发与利用的基本原则。

(1)保护用户隐私是大数据开发的前提

保护隐私是国家和法律对公共基础设施提供者的硬性约束，也是电信行业的基本要求。在大数据开发过程中，采用行业内最高的安全等级存储和处理用户数据，将原始数据对外全方位屏蔽，不会针对个

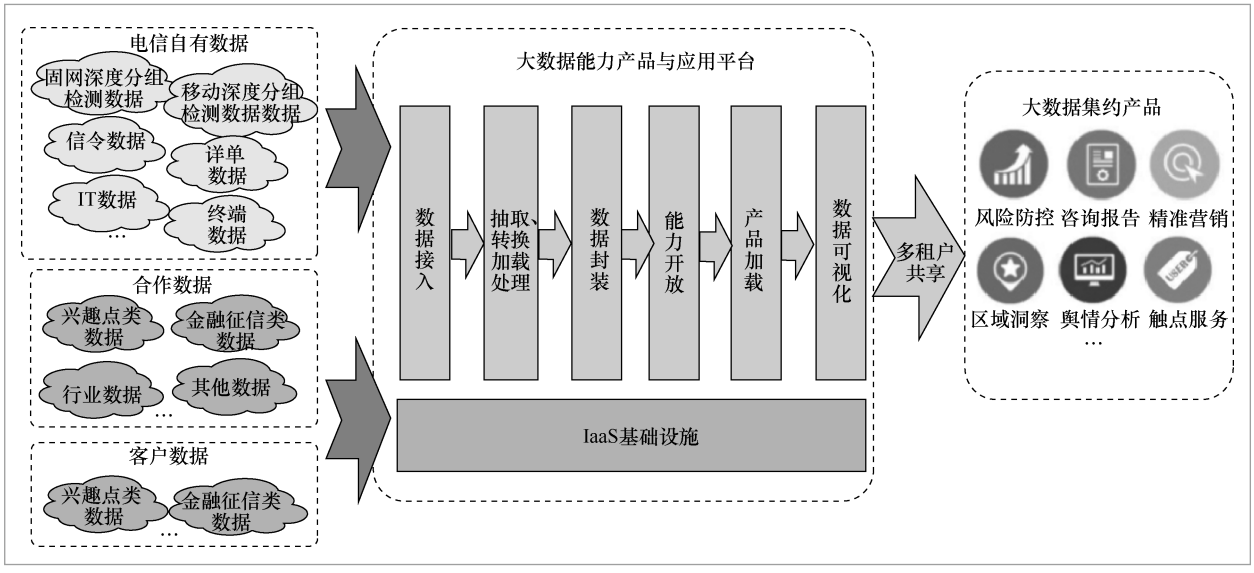


图1 中国电信大数据能力产品与应用体系

体进行分析，所有结果仅进行状态匹配和标签输出，而且所有的数据分析都在中国电信自有平台上进行。

（2）为公众和社会服务是中国电信的理念

作为一个大型央企，广泛促进社会福利水平，保护公众隐私不受侵犯，为企业和个人提供高效数据服务，提升数据社会价值，是中国电信在大数据开发主要考虑的方面。

（3）建设专有的大数据应用平台进行开发

集约地整合、处理、分析所有电信端数据，通过搭建自有服务器资源，保证电信团队、合作企业有足够的资源在电信的平台上做深入的数据分析。能够为企业级应用提供从数据整合，到计算能力、发布平台全流程的服务。

（4）广泛的业务合作，支撑行业内产业链企业的发展

电信大数据的工作任务是提升数据的应用价值，通过打造大数据平台，吸引行业内的各类企业共同开发利用，并与各自的数据资源进行融合。业务定位是提供大数据基础能力支撑，与产业链各方一起促进大数据产业发展，共同成长。

（5）营造健康发展环境，建设大数据应用生态

中国电信致力于建设开放、合作的大数据应用生态。与产业链各方共同营造安全、合规的数据使用环境，有助于未来电信数据能够更好地对外服务。如图2所示，在面向最终客户提供大数据应用服务时，产业链各方充分发挥各自在数据、产品建模、平台技术、解决方案提供方面的作用，实现优势互补，合作共赢。

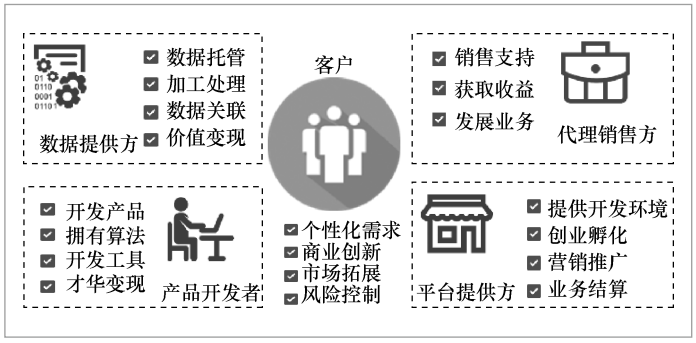


图 2 共建大数据应用生态

4 电信大数据产品和应用

4.1 天翼大数据“4+1”产品体系

中国电信天翼大数据现有“4+1”的产品体系，包括两大类型：数据型产品和平台型产品。

（1）数据型产品

数据型产品主要依托中国电信的数据资源，同时整合外部数据资源（如金融、电商等行业），开展分析、挖掘类数据业务，服务形态主要包括：标签、报告以及SaaS应用。其中，“星图”系列以用户画像和分析为主，分别是风险防控及精准营销2类产品；“鲲鹏”系列以区域分析、群体趋势分析、群体画像为主，分别是咨询报告及区域洞察2类产品。

做数据型产品的目的是更好地从非运营商业务的视角来理解数据，了解数据如何更好地为行业服务，如何有效地与产业链合作伙伴协同。

（2）平台型产品

平台型产品为合作而生。“飞龙”系列大数据云PaaS提供资源托管、数据处理分析、产品孵化3类服务。

大数据离不开云计算基础设施，依托中国电信“8+2+X”的云资源布局，通过构建云计算PaaS平台产品，提供比基础设施层更高、更丰富的平台服务，降低用户使用大数据挖掘门槛，使得开展大数据业务的企业无需担忧技术实现问题，而是将更多

的精力和资源投入对需求的挖掘、分析和满足上；让传统企业能更快、更高效地通过分布式计算框架、完善的数据分析工具组件，实现大数据时代的IT升级换代、同时，通过PaaS平台能力开放以及平台敏捷可靠的开发环境，越来越多的应用开发者、越来越丰富的数据能力为整个产业链提供了有力的生态保证。

4.2 电信大数据产品应用

（1）终端咨询报告

利用中国电信拥有的完整终端自注册信息以及终端用户数据，判断用户终端的使用状态、使用行为特征、消费能力以及偏好等数据，通过数据整合与能力封装，提供终端分布、终端使用行为分析等分析报告服务。

针对终端厂商，提供查询自有品牌终端及竞争伙伴终端的相关数据及趋势分析，分析本产品和竞争产品的市场份额、终端网龄、终端生命周期、换机流向，助力终端设计生产。针对终端销售渠道，提供销售终端份额、终端规模增速、价格构成、价值贡献等信息查询和分析功能，提升销售渠道快速掌握销售市场动向、调整销售策略的能力。针对应用开发商，提供应用渗透率、应用的终端市场占比、应用的使用周期等信息的实时查询，帮助开发商更快、更准确地了解应用市场动向。

（2）精准营销产品

基于运营商多维数据的交织分析，通过关联挖掘海量电信数据和互联网数据，对用户进行标签化处理；与传统互联网标识不同，电信标识体系能更精准识别自然人，通过结合兴趣标签和用户属性标签，更好地服务行业客户；并通过“用户行为—兴趣—产品”的关联标签，结合电信各种新式媒体和触点，将企业营销信息推送到比较准确的受众群体中，为企业节省营销成本，为用户找到合适的需求点，达到ROI（投资回报率）最大化的效果。

（3）区域洞察商业选址

在中国电信的海量数据中，还有一类最有价值的数据就是海量用户的位移数据。依托中国电信移动网络的蜂窝模型及用户的位移，鲲鹏—商业选址产品提供了更有效的数据分析能力，通过海量的用户样本更精准地实现对区域商业价值的评估，改变了传统依托“公开数据+扫街调查”，通过少量样本进行商业选址的传统区域价值评估模式。

依托中国电信大数据，通过用户的区域通信行为，结合POI信息，提供区域常住人口特征分析、车流人流分析、各类商业业态分布和区域竞争信息，让商业选址更智能，真正从行业的视角，以数据的方法帮助客户以最优的性价比选择线下商铺的地址，支撑商铺的运营。

（4）人口流动分析

随着人们生活水平的改善，越来越多的人在节假日选择出行、旅游。公安、旅游等部门都面临如何有效地在人群聚集的状态进行及时的安全监控预警和高效地进行区域人员的分析和预测，避免公共场所群体安全事件发生的问题。

中国电信多样化的数据、政府相关部门的数据、合作伙伴的数据等多源数据，实现了以移动用户的实时数据为基础，有效地对区域人流进行信息化监控预警、分析及服务，让政府的管理机构可以以科学的技术手段实现对关注区域的人流密集程度、流动方向、人流群体的结构、人流特征等多个维度信息的及时掌控。通过实时/准实时的数据汇聚、清洗、分析，各类人流热图的呈现，开发了多种可视化手段。

人口流动类宏观产品已经在流动人口分析、疾病防控、京津冀一体化规划等进行了有益的尝试。在2015年广西东盟

博览会上，中国电信与合作伙伴一起为大会主办方提供了包括互联网专线、IPTV、Wi-Fi及大数据安全预警在内的会展解决方案。通过对手机用户数据、信令位置数据和现场视频数据的关联，运用大数据建模和可视化组件，为展会提供了实时人流监控和人群结构分析服务，做到了及时、准确、可靠的安全预警，有效降低了安保成本和风险。

4.3 电信大数据应用推广

（1）不断深化产业链合作

中国电信始终秉承“合作共赢”的经营理念，在大数据领域不断加强与产业链的开放合作。经过一年多的开发以及与大量厂商的合作开发，电信大数据产品一系列产品已经能够满足海量业务的调用，能够提供高性能的平台运算能力。

2015年11月中国电信正式发布了“天翼大数据”品牌，推出精准营销、风险防控、区域洞察、咨询报告4类数据型产品和大数据云平台型产品，重点服务于旅游、金融、广告、交通、政府等行业和部门。其中，风险防控产品基于中国电信用户标签数据建立用户信用模型，主要服务于银行、保险、征信、P2P等金融机构；区域洞察产品基于中国电信用户位置标签数据，为道路交通、区域人流分析、商业选址分析、智慧城市建设、智慧旅游建设等领域提供数据服 务。

在发布会现场，中国电信与浪潮集团、全联房地产商会、东方国信科技股份有限公司、中诚信征信有限公司、中智诚征信有限公司、华为技术有限公司、中兴通讯股份有限公司、神州泰岳软件股份有限公司等10余家合作伙伴签署了战略合作协议。中国电信将与战略合作伙伴在大数据产品和解决方案等领域持续开展深度合作。

（2）积极推动中国企业大数据联盟（BDU）发展

通过建立数据标准、交换规则，推动跨界合作，创新商业模式，提升参与各方大数据应用的整体水平，提升产业竞争力；汇聚各方力量，吸收国内外先进经验，使联盟成为推动技术进步、应用创新的中坚力量，为大数据产业健康发展做出贡献。

（3）推出大数据成长计划

该计划旨在构建有影响力的大数据生态圈，以中国电信大数据开放平台、高价值数据为支撑，面向行业伙伴提供数据、产品、销售3种合作模式，快速形成聚合效应，促进中国大数据产业健康发展，拉动信息消费，为推动社会转型升级做出贡献。立足于现有平台和未来发展，联合各类企业、科研单位、高校单位等，共同成长。同时中国电信主办了大数据分析竞赛，推动大数据分析在未来人群中的认知、发展和人才储备等。

5 大数据分析案例——电信大数据在政府流动人口分析中的应用

业务需求：分析某省份省会城市辖区中流动人口的比例、构成以及人群的特点，推演出其在医疗卫生方面的需求，为政府和组织的服务提供参考。

将需求分为两个部分，第一部分为如何尽可能准确地描述流动人口，尽可能多地将真实的流动人口提取出来，提高准确率；第二部分为对确认的流动人口进行人口学特征、网络使用偏好、居住信息、活动区域、家庭情况、工作情况方面的分析，支撑政府医疗卫生服务方面的措施推进。

（1）明确流动人口的定义

根据项目的需求，将从省内非省会城市迁徙而来、时间高于1个月的用户设定为研究对象，其中将居住时间超过3个月（可

调）的用户定义为流动人口。分别从时间、位置方面初步区分流动人口群体。

（2）人群初步区分

分析前提：所有“在用”状态的用户，将设定几个用于区分人群的标准，见**表1**，综合如下。

由于不知道户籍方面的信息，单纯从电信数据看，A部分是最有可能产生流动人口的群体；其次为B部分，即流动人口在居住地换本地号码的情况；第C部分需要根据户籍信息是否变动、居住时间等条件判断，根据辅助条件，少量归入流动人口的群体；第D部分为当地居民的可能性更高，认为非流动人口群体。

然而以上分类较粗，错误率会较高，因此加入了其他的辅助筛选条件，如进入本市时长、是否有省内漫游、是否有省内长途电话。

根据就近和信息有效的原则，从开始研究的月份之前倒推6个月开始积累数据，对每月居住时长达到某一阈值的用户，折算为居住一个月。

有省内漫游和省内长途通话的用户将比无省内漫游和省内长途通话的用户为流动人口的可能性更高。

通过以上条件筛选，最终筛选出可能性最大的流动人口的人群，总计约10万人。根据电信用户的比例计算，符合需求的流动人口总量应为70万~80万人。

（3）通过模型进一步扩大流动人口筛选的范围

根据与需求方的深度沟通，在以上筛选方法的基础上，将流动人口与非流动人口进行对比测试，从相关数百个字段中挑选出了30多个最相关变量和衍生变量，将相关变量分为核心变量、辅助变量，并对核心变量进行权重划分。

经过各类模型分析结果对比，选择了人工神经网络作为最终的模型。初步的结果显示，基本上能够将目前数据样本中绝大部分疑似流动人口的用户识别出来，并应用于具体的数据分析工作。

表1 根据人和手机号的归属地划分人群

判断标准	本省非本市居民	本市居民
本省非本市手机号码	A	C
本市手机号码	B	D

（4）部分分析结果举例

通过每月数据的监测，对每月流动人口的变动进行描述，得到了一段时间内人口流动的波动信息和人口的基本信息，如**图3**、**图4**所示。

通过可视化方法，在地图上显示出流动人口的分布、每日流向等信息，还能以动态的方式展示。结合POI等信息，还可以分

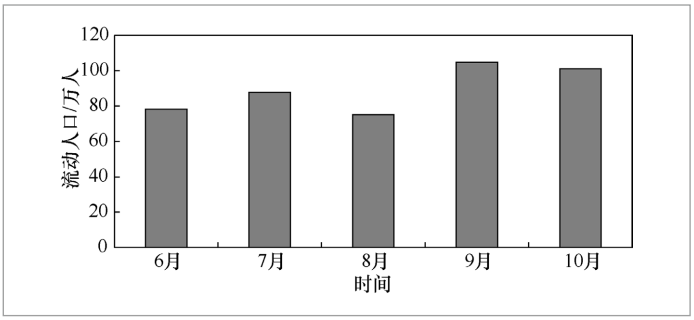


图3 流动人口月数量分布

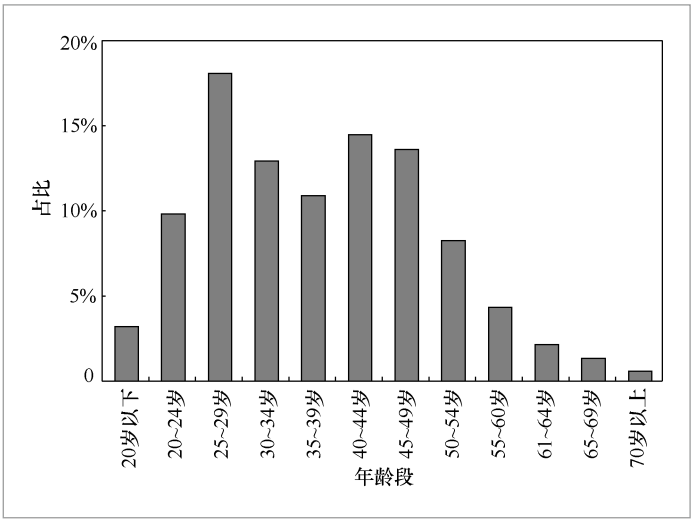


图4 流动人口年龄占比情况比较

析出流动人口生活环境状况等。

利用组合模型，可以分析出特定人群的分布、人群特征、活动特征等信息，例如通过对用户网络行为和位置行为建模分析，能够区分出育龄妇女人群以及她们大体所处的孕育阶段，能够更加精准地为政府决策、公共卫生服务提供参考。

6 结束语

大数据开发的最终目标是行业应用，它依托于大量的数据、强大的分析资源、各类优秀的业务模型以及对垂直领域的洞察。中国电信作为运营商级别的大数据开发者，能够在数据、平台、合作等方面为全社会提供基础资源，促进各行业大数据的开发、融合、应用。

中国电信已经开发了“4+1”的产品体系，并将开发更多的大数据产品、更多的接口，与更多的企业合作。大数据的深度利用，将成为社会经济发展的重要推动力。

参考文献:

[1] 童晓渝，张云勇，房秉毅，等. 大数据时代电信运营商的机遇[J]. 信息通信技术，2013(1):5-9.
TONG X Y, ZHANG Y Y, FANG B Y, et al. Opportunities for Telecom operators in the big data age[J]. Information and Communications Technology, 2013(1):5-9.

[2] 黄勇军，冯明，丁圣勇，等. 电信运营商大数据发展策略探讨[J]. 电信科学，2013，29(3): 6-11.
HUANG Y J, FENG M, DING S Y, et al. Big data development strategy for telecom operators[J]. Telecommunications Science, 2013, 29(3): 6-11.

[3] HORNIK K, STINCHCOMBE M, WHITE H. Multilayer feed forward networks are universal approximators[J]. Neural networks, 1989, 2(5): 359-366. □

第一届大数据科学与工程国际会议（2016）

BDSE
大数据科学与工程国际会议
INTERNATIONAL CONFERENCE ON BIG DATA SCIENCE AND ENGINEERING

THE 1ST INTERNATIONAL CONFERENCE ON BIG DATA SCIENCE AND ENGINEERING (BDSE2016)

时间：5月25-26日 | 地点：中国•贵州•贵阳•万丽酒店

汇集多位院士和国内外大数据领域知名专家学者
聚焦国际大数据技术研究新进展！

以合理
专业的视角解读大数据

以跨界
融合的理念驱动大数据

以历史
发展的眼光认识大数据

科学

创新

以开创
继承的思维培育大数据

认识

人才

大会主办单位：人民邮电出版社、中国计算机学会大数据专家委员会、数博会

承办单位：信通传媒•《大数据》杂志

会议网址：http://bdse2016.j-bigdataresearch.com.cn/　联系方式：010-81055475/5448/5490　电子邮箱：bdse2016@bjxintong.com.cn

ISSN 2096-0271

邮发代号：2-537　国外代号：C9118　定价：35.00元

9 772096 027162 05>