电信大数据关键技术挑战

曾嘉[1]，刘诗凯[2]，袁明轩[1]
1. 华为诺亚方舟实验室，香港 999077；2. 华为大数据分析产品部，江苏 南京 210012

摘要
大数据逐渐对用户体验和生产效率带来颠覆性影响。电信大数据来源于运营商通信网络平台的BSS和OSS，沉淀了海量用户7个维度的信息：1维用户真实ID、1维行为数据、1维社交数据、1维时间数据和3维空间数据。运营商构建电信大数据分析平台，通过对7维用户数据建模，可以实现3个数据业务方向的升级：用户洞察、网络洞察和数据开放。着重探讨电信大数据分析平台遇到的9个关键技术挑战和可能的技术突破方向。

关键词
电信大数据；用户洞察；网络洞察；数据开放

中图分类号：TP 391　　文献标识码：A　　doi: 10.11959/j.issn.2096-0271.2016034

Key technical challenges in telecom big data

ZENG Jia[1], LIU Shikai[2], YUAN Mingxuan[1]
1. Huawei Noah's Ark Lab, Hong Kong 999077, China
2. Huawei Big Data Product and Solution, Nanjing 210012, China

Abstract
Big data has been improving steadily user experience and productivity. Telecom big data comes from the telecommunication platform composed of the BSS (business supporting system) and OSS (operation supporting system), which accumulate billions of customers' 7-dimensional (7D) data including 1D for real ID, 1D for customer behavior data, 1D for social network, 1D for time series and 3D for spatial information. Telecom big data platform can support modeling of 7D customer data, which enables three business upgrades, including customer insight, network insight and data openness. 9 technical challenges of telecom big data analytics and possible solutions were described and discussed.

Key words
telecom big data, customer insight, network insight, data openness

1 引言

大数据的3V特性(volume、variety、velocity)正在逐步改善用户体验和生产效率。电信运营商提供基础通信平台连接每一位用户。每天数以亿计的用户在管道中留下的通信行为数据构成电信大数据。如何挖掘大数据来智能支撑运营商各项业务并进行业务转型,是一个需要深入思考的问题。首要任务就是汇集丰富的用户行为数据,存储在统一的电信大数据分析平台,并根据业务需求灵活部署统计、规则和预测算法,在不改变业务人员使用习惯的情况下做到及时、多屏、准确、直白、客观地沉淀用户数据(例如不同时间跨度的知识标签),帮助企业减少营销成本的同时做到360度营销,降低业务人员的学习成本,赋能、提高营销效率。总体而言,电信大数据沉淀了海量用户7个维度的信息:1维用户真实ID(基本信息)、1维行为数据(通信行为、互联网行为、消费行为、投诉行为、网络体验、反馈行为)、1维社交数据、1维时间数据和3维空间数据(室外宏基站定位和室内微基站定位)。通过对7维用户数据建模,可以实现运营商在3个数据业务方向的升级:用户洞察、网络洞察和数据开放。如**图1**所示,电信大数据平台的目标是实现用户、网络和数据的统一自动化管理,实现“三个了解”和“三个提升”,即了解用户、了解产品(服务和渠道)、了解网络,提升营销转化率、提升决策准确率和提升自动化率(数据化→信息化→智能化)。

全生命周期的用户洞察是电信大数据的基础,目标是改善用户体验,提升营销效率,从而沉淀更多用户的行为数据作为反馈。以用户为中心的网络洞察有助于提升运营商在网络规划、网络建设、网络优化、网络维护方面的投资效率,改善用户网络体验,并降低运营成本。面向全行业的数据开放的重点是利用电信大数据优势构建数据产业生态链,使其能提供面向全行业的数据服务,例如帮助行业客户进行获客、营销、选址分析、人流量检测、区域价值规划等。然而,在3个业务方向的升级都急需强有力的电信大数据分析平台支撑,这将面临9个方面的技术挑战,分别是特征工程、预测算法、根本原因分析、实时分析、时空数据挖掘、知识管理、多媒体

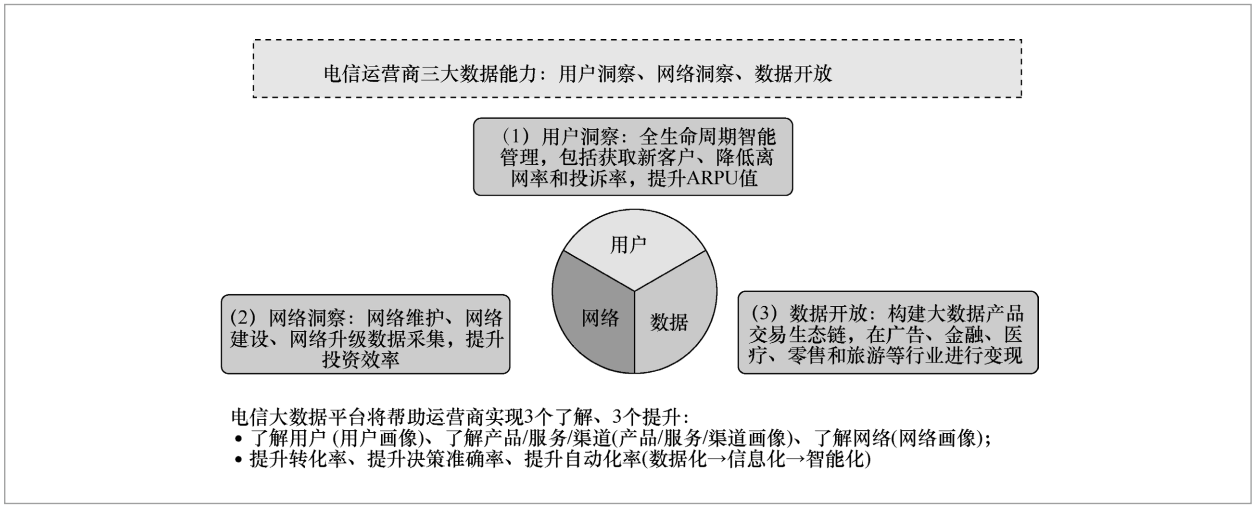


图1 电信大数据平台的目标

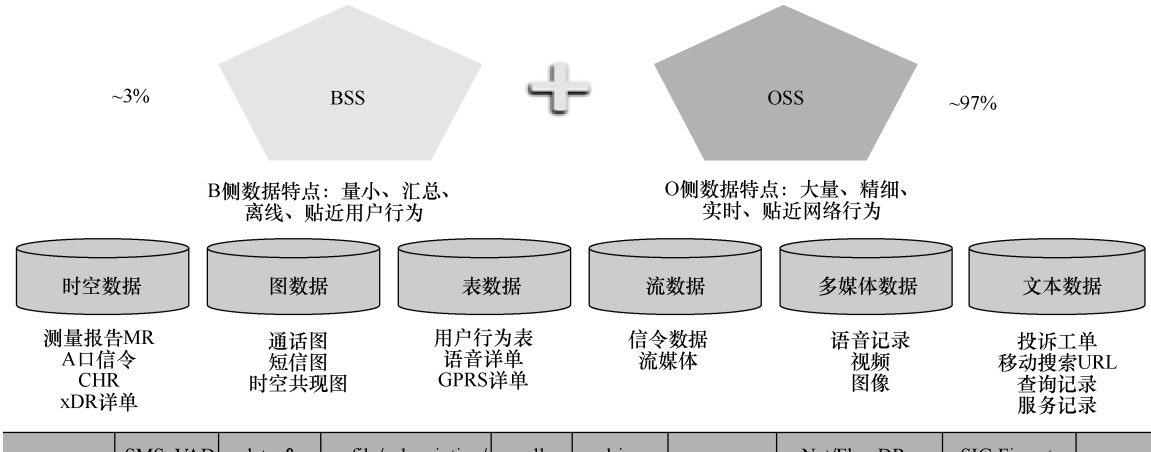
数据挖掘、图数据挖掘和隐私保护。本文重点描述这些技术挑战的来源和可能的解决方案，讨论如何构建高效的电信大数据平台。

2 电信大数据分析

电信大数据平台是一个提供统一数据存储、分析处理、数据服务的云计算平台。如**图2**所示，平台需要融合并存储来自BSS（B域）和OSS（O域）系统的数据，提供6种通用数据的处理。BSS是商业支撑系统，包括客户关系、计费、营销和传统商业智能系统，记录用户的话单、账单、基础信息和营销反馈记录。BSS数据特点是量小（约占电信大数据总量的3%）、汇总、离线（非实时更新）、贴近用户行为。OSS是网络运营支撑系统，包括基站、传输、固网和核心网等网络单元（CS系统负责语音/短信、PS系统负责上网流量），记录大量信令数据，包括用户联接网络体验、互联网内容和位置信息。OSS数据特点是量大（约占电信大数据总量97%，主要是位置数据和互联网内容数据）、精细、实时和贴近网络行为。以600万个活跃用户为例，每天产生大约14 TB数据，这些数据大部分来自OSS，通过扩展，可估算中国12亿用户每天产生的数据量。尽管数据来源不同，但所有数据可抽象为六大通用数据类型，分别是时空数据、图数据、表数据、流数据、多媒体数据和文本数据。建模算法可以面向这六大类通用数据类型进行设计和部署。

电信大数据核心资产是海量用户的7个维度信息。如**图3**所示，这些数据可以支撑对内和对外服务优化，如全生命周期的用户管理和基于位置的服务。用户洞察的核心是围绕用户回答7个W的问题：who（用户ID）、when（时间）、where（空间）、what（行为结果）、how（行为过程）、why（行为根本原因）、Web（社交）。通过7D用户数据建模实现7W洞察是电信大数据分析的核心需求。

大数据时代，建模思维逐渐从研究各种映射算法到研究数据本身，如从丰富的数据中抽取更加合理的特征表示、从



数据源	SMS, VAD records	data & voice CDR	profile/subscription/ account Info	call center	drive test data	MR data	Net/Flux DR、SC data	SIG Fic net data set	total
数据量/天	15 GB	26 GB	94 GB	0.14 GB	0.02 GB	1 900 GB	2 880 GB	9 000 GB	13.8 TB/天

上述是大约600万活跃用户每天产生的数据量，可以扩展估算全中国12亿用户每天产生的数据量

图 2 融合 BSS 和 OSS 数据

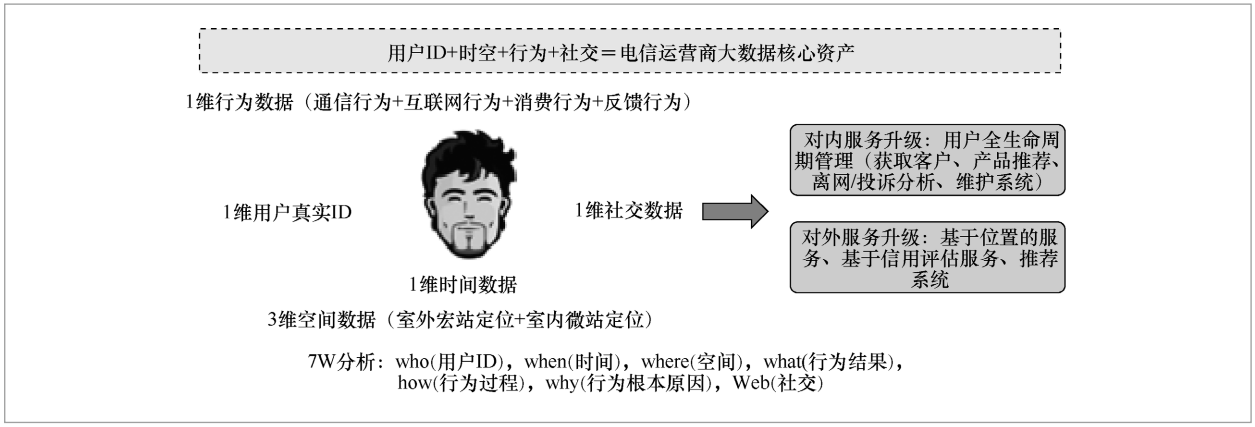


图 3 7 维度用户数据建模和 7W 洞察

数据中根据业务规则自动定义标签和训练数据以及利用用户营销反馈数据，自动化整个营销流程等。从数据出发，寻求合适、好用的算法是建模的核心。清晰定义训练数据，熟悉业务流程，才能将挖掘算法的价值发挥到最大。用户和基站联接行为将BSS和OSS数据打通，因此通用的用户模型可以用二分图表示，即一层节点是用户，另一层节点是网络，两层节点之间的边表示用户联接网络的时间。总体而言，可以通过电信大数据的二分图表示实现用户洞察、网络洞察和数据开放3个业务目标。

3 技术挑战

电信大数据分析面临9个关键的挑战，分别是特征工程、预测算法、根本原因分析、实时分析、时空数据挖掘、知识工程、多媒体挖掘、图挖掘和隐私保护。

3.1 以时空数据为核心的特征工程

随着移动设备和移动互联网的普及，随时随地使用移动终端已经成为人们的一种基本生活习惯。因而电信数据成为获取城市用户、区域细粒度时空行为信息的重要数据源。这些细粒度行为信息可以被用作建模的重要特征，从而大幅提升电信数据挖掘效果[1]。因为电信数据来自多个数据源，如BSS（B域）的数据来自CRM（customer relationship management，客户关系管理）、账单、BI（business intelligence，商业智能）、客服和渠道等系统，OSS（O域）的数据来自于MR（measurement report，测量报告）、Gn口和Mc口等系统，时空和用户ID关联是把这些数据整合成统一特征集合的关键因素。以时空数据为核心的特征工程需要结合B域和O域进行关联分析，找出网络和用户特征的关联性。如**图4**所示，复杂的特征工程可以在以时空数据为核心的各种数据类型上构建。如人的社交关系可以表述为电话网络、短信网络和接触网络（两个人在相近时间、相近地点出现算是一次有效接触）。每个电话、短信或接触都有发生的时间和地点。需要设计有效的算法研究如何在这种有时空约束的图中提取反映用户复杂社会关系的特征。另一个例子是将用户的账单、影响力或者离网行为映射到每个位置上来评估每个位置的价值，从而可以得到高价值用户或者离网用户聚集的位置，开展基于位置的服务和营销。同时也可

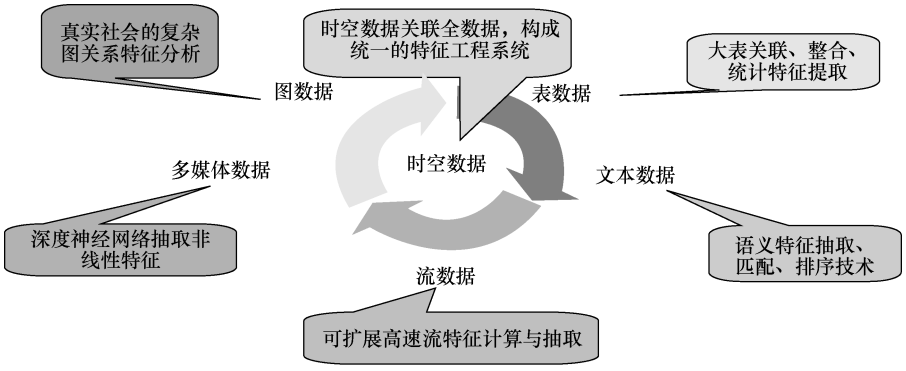


图 4 以时空数据为核心的特征工程

以构建基于用户行为的基站投资分析，得出精确基站选址目标。将通信行为、互联网行为、消费行为、营销反馈行为映射到时空位置坐标，也可以开放给其他行业，输出专业性的评估报告，有助零售业或者旅游业掌握移动用户的行为。例如，西班牙电信Telefonica的Smart Steps洞察方案将时空数据脱敏后开放给行业客户，每年有数千万欧元营收。其他的多媒体数据（客服语音记录）、文本数据（服务记录和移动搜索记录）和用户轨迹数据等都需要设计有效的特征提取算法。所以电信大数据挖掘的第一个核心挑战是以时空数据为核心的异构特征工程。

3.2 更加精准的预测算法

当特征工程完成以后，下一步需要做的是预测（如离网预测[1]），并根据预测结果做决策。业务价值通常取决于预测的精度，精度越高越好。图5对比了传统数据挖掘的预测流程和大数据下的预测流程，主

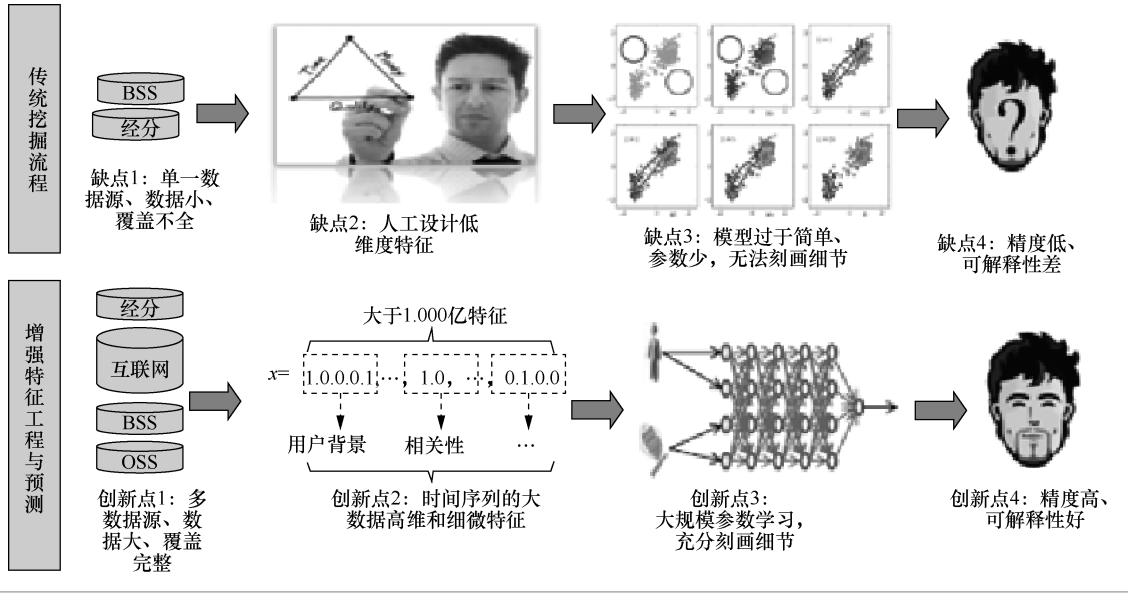


图 5 精确的预测算法

要区别之一是传统数据挖掘采用的预测模型较简单(参数少),无法刻画数据统计分布的细节,而大数据背景下,通常采用大规模参数学习(如支撑十亿级别的模型参数处理百万级别的稠密连续特征向量),从而充分刻画统计细节和数据的相互依赖关系,达到更高的预测精度和更好的解释。传统的梯度下降(上升)算法在并行架构下可以优化大规模参数的神经网络模型,是未来高精度预测算法发展的主要方向之一。但是,电信领域的数据种类相对互联网领域数据种类较少,通常使用较少的特征也能带来业务性能的提升。未来需要更多的实验验证高维丰富的特征对电信业务的有效性和必要性。

3.3 根本原因推断辅助商业决策

商业智能的核心竞争力是分析用户行为的根本原因,即明确哪些主要变量影响用户最终的行为。如图6所示,运营商关心的是何种原因(如网络质量体验、资费、服务体验不好)导致用户离网行为,从而可以针对性地改进产品和服务,持续改善在网用户体验。未来个性化的营销也需要对用户多个行为变量进行排序,从而匹配到最为需要的产品。根本原因推断技术的主要方向仍然是特征变量的排序和变量之间相关性分析。由于大部分特征变量仅仅反映表象,根本原因分析需要对隐藏变量进行推断,然而目前大部分统计学习算法仍然难以有效地产生可以解释的隐藏变量,需要持续探索。

3.4 实时分析能力逐步成为基本需求

OSS数据的一个特点是更新速度快,如用户对网络的体验、网络故障诊断和位置更新信息,都是在秒级或者分钟级采集产生的。这些数据一旦不及时处理分析,将失去商业价值。例如,客户当时上网体验不好(例如无法使用支付宝或者微信支付出租车费),很有可能会即时拨打投诉电话,因此需要即时得到分析结果,并做一些补偿措施,给用户良好的体验。又例如网络故障诊断,需要在故障发生之后,立即分析并隔离相关的网络单元,启动备用方案。基于位置的营销需要及时判断用户的当前位置,推送附近商铺的合理产品,或者是当用户靠近营业厅附近时,推送合适的业务服务。实时分析能力需要流处理架构和在线学习算法,通过统计、预测一个短时间窗口内的数据流,迅速更新模型参数,并做出决策。之后的决策都基于模型,不需要重新学习历史数据,因此大大加快了模型的更新速度和分析速度,达到实时处理的目标。尤其对于海量OSS数据,流处理和在线学习技术是非常必要的。另一个挑战的技术方向是时间序列的挖掘,如何在数据流中快速捕捉数据在时序上的依赖关系(上下文关系),做出准确的预测,仍然十分困难,需要持续研究。

3.5 时空数据挖掘

电信数据相较于其他数据的一个核心优势就是含有用户细粒度的时空行为信息。有效的挖掘并利用这些时空数据可以充分地发挥电信数据的价值。但是,电信

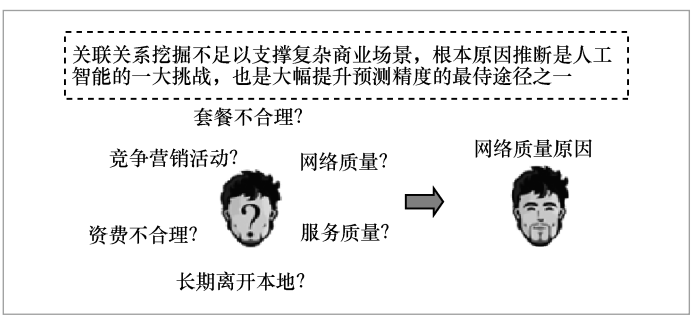


图 6 根本原因推断算法辅助商业智能

时空数据的挖掘面临4个核心挑战：数据的不精确性、数据的超稀疏性、数据的强依赖性和异构性，如**图7**所示。数据的不精确性是指通过电信数据计算的用户位置精度远低于GPS精度（但是电信位置数据的好处是不需要客户端任何负担，位置数据天然存在于网络侧）。如**图7**中浅色圆圈是某区域用户真实GPS位置，浅色圆圈是使用基于距离的定位算法从电信记录恢复的用户位置[2]。可以看到直接使用简单的基于位置的定位算法，数据存在很大的不精确性。如何设计更好的定位算法，如充分利用指纹和地图等信息，降低位置数据的不确定性，是第1个挑战。第2个挑战是数据的超稀疏性。每个用户只会出现在城市的一个很小的区域和一些小的时间片段中。如果把所有用户的时空数据放在一起，把每个小时间片段和地点的组合看成一个记录点，一个用户在绝大多数的记录点都是没有信息的。所以时空数据是一个超稀疏的数据集，如何处理并清洗这种超稀疏的数据集是一个技术挑战。时空数据有很强的时间和空间关联关系，如果按照时间切片或者地点切分将时空数据输入数据挖掘模型，这种关联关系就无法被有效地使用[3]。如何有效地组合使用有效的算法，如时间序列和神经网络来有效地表述时空数据的时空强依赖性，是第3个技术挑战。第4个技术挑战是时空数据和其他数据结合时导致的数据异构性，如图、文本挖掘都需要考虑相关数据产生的时间和地点才能进行更有效的信息提取。

3.6 知识管理是智慧延展的基础

运营商每年有大量业务人员沉淀经验知识用于营销、网络优化和客服。大部分知识都是通过文本的形式保存下来，但是这并不方便查询和寻找知识之间的关系。文本挖掘的一个重要方向是自动构建知识图谱，通过发现文档中知识单元之间的相互关系，方便用户查询和学习。如**图8**所示，左边是从几十万份网络故障相关的文本中提出的关键词（知识单元）和相互之间的关系，通过点击相关的知识单元，可以查到对应的文本摘要，大大缩短定位问题的时间。类似

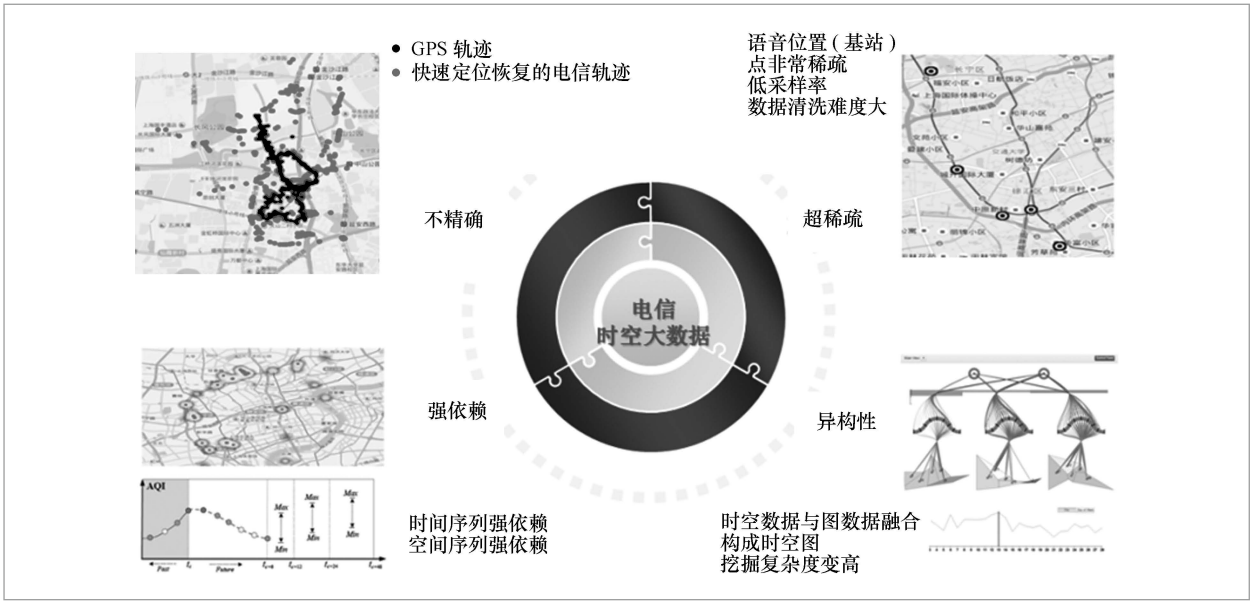


图 7 电信时空数据挖掘的 4 个核心挑战

地，客服系统每年都存有大量用户投诉咨询记录，通过构建知识图谱，可以容易地发现投诉热点，并做出持续改进，节省大量人力、物力。

3.7 多媒体数据挖掘

电信数据中的多媒体数据主要指客服的语音记录。语音记录中包含了客户关注的问题和客服服务质量和有效性信息。与客服人员手工记录的文字信息相比，客服语音信息包含更原始和真实的信息，如客户的情绪、关注点和客服的效率等信息。有效地挖掘这些信息可以自动化地发现资费、网络、服务和竞争对手的问题，提升服务质量。语音数据中提取的特征也可以被有效地应用于其他数据挖掘模型。语音数据的处理包含两个部分，语音识别和文本自然语言处理。语音识别主要有两个挑战，一个是当前电信记录系统很多是8音道数据，数据质量较差；另一个挑战是语音中含有很多方言和电信业务相关专用词汇，需要特殊的算法提升识别精度。语音识别为文本后，需要自然语言处理算法准确地提取关注主题、客户情绪和服务质量等信息。当前深度学习技术已经在百度和谷歌等公司的语音识别[①]和自然语言处理[②]方面显示出强大的能力。如图9所示，如何针对电信语音数据设计合适的深度学习算法是多媒体数据挖掘的技术挑战。

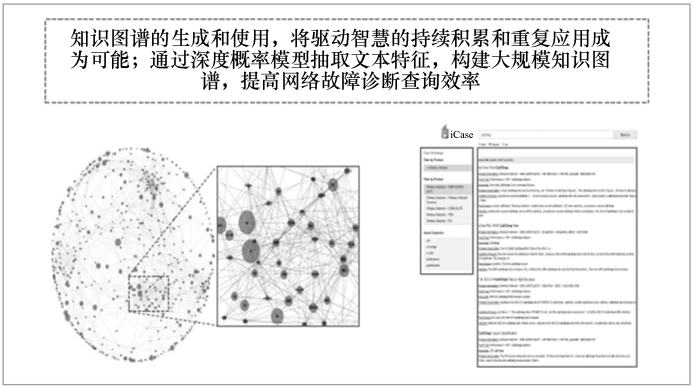


图 8 知识图谱用于沉淀并管理业务知识

3.8 图数据挖掘与社交分析

电信数据包含3种基本的用户社交网络：电话网络、短信网络和用户接触网络。图挖掘技术已经在很多数据挖掘场景（如推荐系

① Baidu. Deep speech[EB/OL]. (2014-12-19) [2015-11-10]. http://36kr.com/p/217970.html

② Google.com. NLP Group[EB/OL]. [2015-11-10]. http://research.google.com/pubs/NaturalLanguageProcessing.html

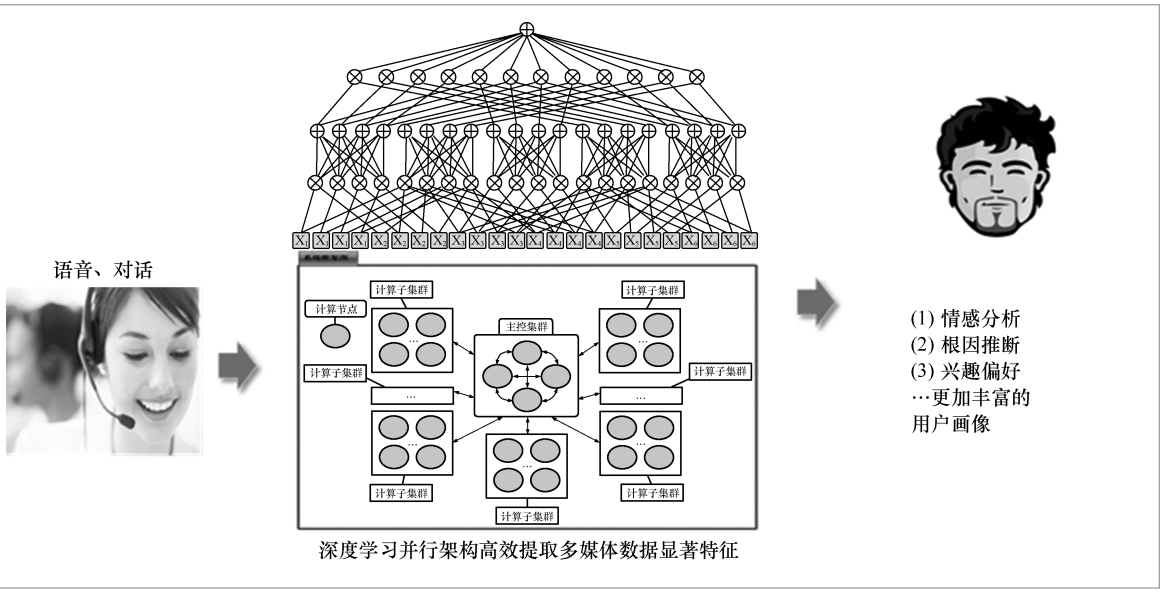


图 9 深度学习技术应用于多媒体数据挖掘

统）中取得了很好的效果。电信图数据的主要特点是每个图不仅包含社交信息，还包含社交发生的时间和地点信息。含有时间和地点的图的分析算法需要新的设计[4]。如何在时间和地点都有约束的网络中构建有效的并行分析算法，并将算法结果转化为模型分析的重要特征是一个技术挑战。如图10所示。

3.9 隐私保护

电信数据含有用户的通信行为、消费行为、互联网行为、社交行为和时空行为等高隐私信息。隐私保护是需要考虑的一个核心技术。当前隐私保护最有效的方法是差分隐私保护[5]。差分隐私将数据分析人员和分析数据隔离，保证攻击者在有任何背景知识的情况下，都只能以极小的概率区分某个特定用户是否在数据集中。如何将差分隐私保护紧密地结合在电信挖掘的算法中是一个值得研究的课题。从当前实际系统需求分析，另外一个更加重要的隐私问题是防止数据滥用技术的研发。当前数据挖掘都是经过用户授权使用数据，但是电信运营商为了保障数据隐私安全，要求分析人员只能在严格控制的内网分析匿名数据，从而隔绝分析人员和分析数据。而推荐系统等应用需要不断迭代的特征工程以保证最优的挖掘效果，在这种场景下的分析技术尚不成熟，例如无法不断迭代特征工程来保证推荐系统等应用的最优挖掘效果。实际商业中最紧迫的场景是和第三方合作，即授权第三方使用数据完成某项数据挖掘任务（用户授权情况下）时，如何限制分享的数据只能被用在这个特定的数据挖掘任务而不被使用在任何其他场景，即阅后即焚的功能。

4 结束语

电信大数据沉淀于通信管道内，覆盖12亿中国用户，需要运营商、设备商和大数据产业链共同努力以发挥其巨大的商业价值。本文提出的9个技术挑战中，一部分已经有相对完善的解决方案，但大部分还需要研发人员和市场人员的努力，在数据挖掘和商业模式方面做进一步突破。2014年是中国电信大数据元年，到2015年，电信大数据已经在用户洞察、网络洞察和数据开放3个业务方向上积累了不少成功的经验。随着技术进步，电信大数据将逐渐释放巨大的商业价值，提升用户体验，降低运

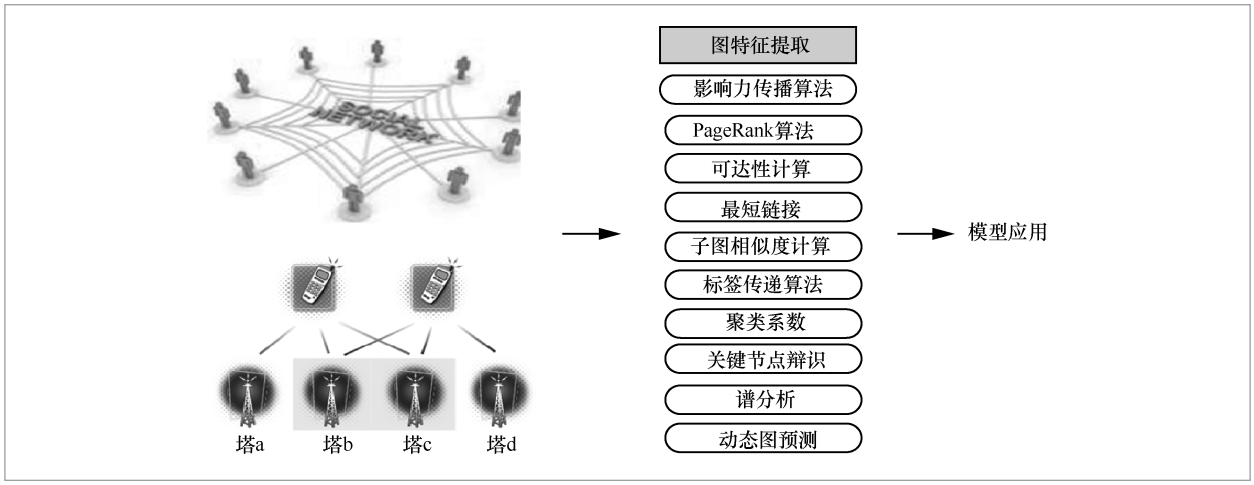


图 10 电信图数据挖掘与社交分析

营成本，催熟整个大数据产业链。

参考文献:

[1] HUANG Y Q, ZHU F Z, YUAN M X, et al. Telco churn prediction with big data[C]// The 2015 ACM SIGMOD International Conference on Management of Data, May 31–June 4, 2015, Melbourne, VIC, Australia. New York: ACM Press, 2015: 607–618.

[2] LI Z T, LI R F, WEI Y H, et al. Survey of localization techniques in wireless sensor networks[J]. Information Technology Journal, 2010, 99(8): 1754–1757.

[3] SHANG J B, ZHENG Y, TONG W Z, et al. Inferring gas consumption and pollution emission of vehicles throughout a city[C]// The 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, August 24–27, 2014, New York, USA. New York: ACM Press, 2014: 1027–1036.

[4] WU H H, CHENG J, HUANG S L, et al. Path problems in temporal graphs[C]// The VLDB Endowment, September 1–5, Hangzhou, China. New York: ACM Press, 2014: 721–732.

[5] HU X Y, YUAN M X, YAO J G, et al. Differential privacy in telco big data platform[C]//The 41st International Conference on Very Large Data Bases, August 31–September 4, Hawaii, USA. [S.l.:s.n.], 2015: 1692–1703.

作者简介

曾嘉（1980–），男，博士，华为诺亚方舟实验室高级研究员和项目经理，主要研究方向为机器学习算法和时空数据挖掘，近期特别在大规模概率主题建模算法做出一系列改进。在相关顶级学术期刊和会议（TPAMI、JMLR、TKDE、TIST、TFS、SIGMOD、VLDB、WWW、ICDM）发表过多篇文章，目前是CCF/ACM会员、IEEE高级会员。

刘诗凯（1983–），男，华为大数据分析产品部主任工程师，主要研究方向为电信业务场景下分析技术的自动化，包含特征表达、参数搜索等。在2015年中国大数据技术大会、2014中国移动技术大会上做过技术专题介绍。

袁明轩（1980–），男，华为诺亚方舟实验室研究员，主要研究方向为电信数据管理与挖掘、时空数据管理与挖掘。2013–2015年，作为核心成员成功完成电信领域多个数据挖掘系统的研发与实际部署应用。

收稿日期：2015–12–30
* 本文为2015中国大数据技术大会（BDTC）演讲约稿