



运营商存量经营大数据平台及其关键技术研究

“基于大数据的互联网化存量经营”项目组,“基于用户感知的运维转型”项目组

(中国联合网络通信集团有限公司上海分公司 上海 200122)

摘要:为了应对运营商从增量经营向存量经营转变的必然趋势,提出做好存量经营的核心是大数据精准洞察,包括大数据平台建设和数据建模两个重要环节。提出了运营商存量经营大数据平台,其可以在主要应用场景下解决运营商当前应用系统普遍存在的问题,并在中国联合网络通信集团有限公司上海分公司(以下简称上海联通)存量经营创新实践中取得较好的商业成果。

关键词:存量经营;大数据平台;数据建模

doi: 10.3969/j.issn.1000-0801.2014.06.019

Research on Big Data Platform and Key Technologies for Inventory Management of Carriers

Project Team of Big Data for Inventory Management,

Project Team of User Aware Network Operation and Maintenance

(Shanghai Branch of China United Network Communications Co., Ltd., Shanghai 200122, China)

Abstract: Carriers need to shift from incremental to inventory management. It was presented that big data platform and key technologies played the important roles in inventory management. The proposed technologies could solve the existing problems in current carrier's supporting systems, and better business results were achieved in Shanghai Unicom business practice.

Key words: inventory management, big data platform, data modeling

1 引言

数据业务的爆炸式增长并未给运营商带来收入的飞速增长,两者差距扩大的趋势日益明显。同时,各类OTT应用也动摇了运营商短信和语音收入的基础。新用户数量增长乏力,而存量用户10多倍于新用户,从而增量经营体系向新型存量经营体系转变是运营商现实的选择。存量经营的理念是以大数据为驱动引擎,使整体存量经营向精准营销、个性化服务转变。运营商大数据包含用户、终端、业务、网络、经营、客服等全量数据,具有巨大的商业潜力。运营商大数据存量经营是通过深度挖掘大数据价值,

支撑用户维系挽留(以下简称维挽)、通信价值提升和非通信价值变现等商业场景,这些都需要运营商先进的大数据平台以及建模技术作为基础。首先,通过大数据平台将用户、产品、渠道和营销4个关键要素融为一体,实现运营从“粗放型”向“精细化”的过渡;其次,通过大数据建模实现用户洞察和全生命周期价值,基于用户偏好匹配业务资源,并利用互联网化的渠道体系实现场景化接触与干预,按照用户个性化需求进行营销活动。

2 存量经营的商业诉求

存量经营商业价值的成功是运营商大数据创新的最大

诉求。存量经营的商业场景主要包括:对内支撑用户维挽、通信价值提升等运营活动;开展对外数据合作,实现非通信价值变现,应用于广告投放、零售选址、交通规划等行业领域。图1显示了运营商存量经营的3类商业场景,具体介绍如下。

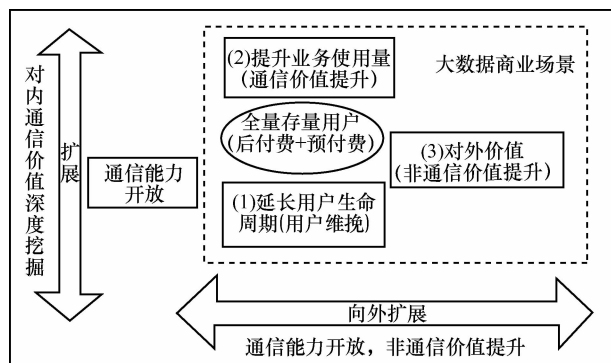


图1 运营商存量经营商业场景主要应用分类

- 用户维挽:通过大数据建模精准预测潜在离网用户,分析用户的离网成本,进行个性化维挽,可以实现用户维挽与成本资源配置的协调,将资源向高价值、高离网概率用户倾斜,提升存量用户数及业务收入的保有率。
- 通信价值提升:长期以来,运营商缺少对用户潜在价值和社会传播价值的二次开发。把握存量经营的价值规律,尽早实现存量用户潜在消费能力的释放,不仅可以实现收入倍增,而且可以带来良好的社会品牌传播效益,助推经营局面的进一步优化。通过大数据分析挖掘用户终端、套餐、偏好、位置等数据信息,可以针对不同用户个性化营销通信叠加包、定向包等产品,释放用户潜在价值。
- 非通信价值变现:通过大数据平台聚合用户数据,经过脱敏加密将数据封装成标准产品服务,面向行业合作伙伴开展业务,提供商业数据分析服务、广告投放支撑、移动支付系统,实现运营商大数据资产的非通信价值变现,前景十分广阔。

3 运营商数据管理现状

存量经营的核心是对大数据的管理分析和挖掘,如何实现商业价值。运营商组织结构复杂,产品和业务种类繁多,数据量巨大,数据管理和挖掘的挑战很大。尽管运营商拥有强大而完备的IT系统能力,但受到运营商传统“重计费、轻管理”技术思维的影响,除了计费方面的能力比较突

出外,运营商在大数据的商业价值挖掘方面存在很大不足。在大数据时代,运营商数据管理的机制和能力还很薄弱,客观制约了各种数据潜在价值的挖掘。运营商当前数据处理系统的主要问题为:缺乏统一的大数据平台,未能统一进行数据管理,包括数据采集、存储、共享和处理等;缺乏有效的数据建模技术,未能充分发掘大数据商业价值。

(1)缺乏统一的大数据管理平台

首先,运营商主要的数据源有:O域的CS/PS信令数据、网管数据、位置数据等;B域的计费、营账、客服、增值业务数据等。采集渠道多,通常分级、分地区、分系统建设,投资重复,整体规划不足,数据标准化程度低,指标口径不一,字段含义不一,无法形成有效的数据资产。

其次,在运营商当前的垂直体系架构中,每一个数据应用系统都是孤立建设,导致不同系统间的数据难以共享、应用上线慢。

最后,电信运营商建有以数据仓库为核心的经营分析系统,通常采用小型机+高性能存储架构进行建设,针对传统话单日志等结构化数据设计,不具备非结构化数据与流数据的分析处理能力。

(2)缺乏大数据建模技术来发掘数据的商业价值

大数据建模需要高效的机器学习和数据挖掘算法,并根据业务场景合理地对数据进行分析处理。首先,运营商数据建模手段单一,主要靠人工设定简单规则实现,很少利用丰富的机器学习算法自动分析数据的潜在价值;其次,运营商不清楚如何利用丰富的大数据实现业务目标,缺乏既了解运营商业务,又了解大数据挖掘技术的跨领域人才。

4 大数据管理平台架构与功能设计

大数据平台的总体目标是构建统一的数据采集、存储、挖掘与分析处理能力。平台必须支持海量异构数据源的采集,包括运营商内部的结构化、半结构化、非结构化数据和外部第三方数据,并解决数据量暴增所带来的存储与计算性能问题。

大数据平台选用目前主流的Hadoop分布式系统。基于Hadoop的大数据平台不仅是底层分散的各种数据源的汇聚平台,更重要的是要在数据汇聚基础上对数据进行有效跨域整合,并基于强大的数据挖掘分析模型,对上层应用提供各种跨域数据分析能力,从而使应用能专注于自身的业务逻辑,快速创新,促进上层业务应用百花齐放。大数据平台的关键特点如下。



- 基于 Hadoop 的 x86 服务器集群, 支持强大的可扩展能力, 可以随数据量增长而平滑扩容; 将 Hadoop 与虚拟化云技术结合, 可以实现更灵活的分布式资源管理能力; 基于 Hadoop 流处理技术, 提供实时数据分布式处理能力。
- 数据统一采集、存储、整合、共享。基于数据透明访问模块提供数据统一访问管理, 实现应用与数据解耦, 同时增强数据安全管控能力, 提升用户隐私保护能力。
- 整合数据挖掘模型, 构建运营商统一的商业洞察与网络洞察组件或模型库, 提供统一的能力服务, 快速部署应用。

基于 Hadoop 的大数据平台系统架构如图 2 所示。架构功能逻辑上主要包括三大块: 数据层、能力层、管理域, 具体介绍如下。

数据层主要包括分布式 ETL 模块、数据库、数据透明访问模块。

能力层主要包括大数据挖掘建模基础能力组件、商业洞察组件、网络洞察组件、能力总线等模块。其中, 大数据挖掘建模基础能力组件是大数据的加速器, 主要提供数据拟合、聚类、机器学习等核心算法库, 用于寻找数据间的关联关系; 也包括文本分析、语音分析、视频分析、图分析、自然语言处理、搜索引擎等各种类型的数据分析处理技术;

还包括数据分析集成开发环境。能力层是整个大数据的核心, 使大数据平台区别于传统数据中心, 真正具备数据智能。能力层的大数据建模技术将在第 5 节进行详细介绍。

管理域主要包括系统管理与数据治理, 系统管理主要管理软硬件资源, 提供简单易用的系统操作维护界面, 包括集群安装、部署管理、软件升级管理、节点管理、服务管理、任务管理、配置管理、集群监控、告警管理、日志管理等功能; 数据治理主要管理系统内存储的数据, 包括元数据管理、数据质量管理、数据生命周期管理、数据安全等。

4.1 数据层

数据层的核心是数据整合和数据透明访问。数据整合主要采用统一任务调度控制所有 ETL 数据采集模块。基于元数据配置执行数据采集流程, 完成数据采集、清洗、转换、关联、入库等各种操作。通过外部数据源与大数据平台数据库之间的数据传输、大数据平台内库与库之间的数据传输, 实现对整个系统数据处理逻辑的全局调度。使用 Hadoop 的 MapReduce 技术, 分布式并行执行各种操作, 整个采集性能非常高, 较之传统 ETL 算法, 性能提高 10 倍以上。数据整合模块的框架如图 3 所示。

基于每种数据库的特点, 可以灵活配置数据存储策略, 优化系统性能。Hadoop 存储 ODS 操作型数据和非结构化用户详单数据, 如用户上网日志、用户详单原始数据;

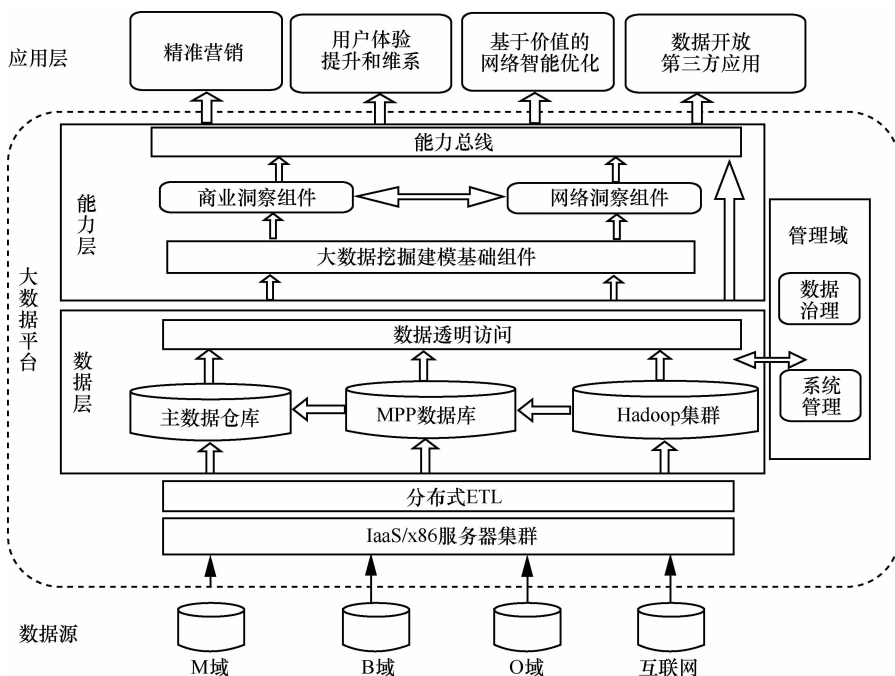


图 2 大数据平台系统架构

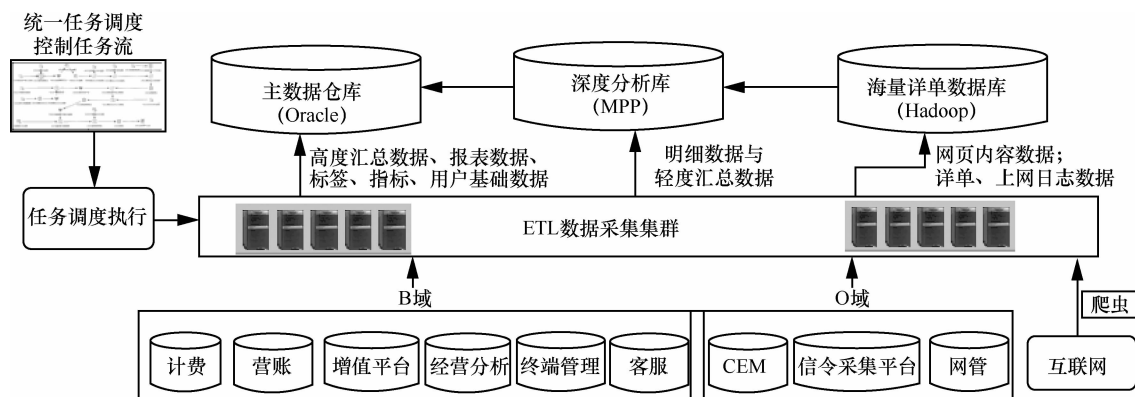


图 3 数据整合模块框架

MPP 数据库存储明细数据与轻度汇总数据;主数据仓库存储高度汇总数据、报表数据、多维数据、标签库、指标库等结果数据。

数据透明访问模块提供统一的数据访问接口,屏蔽底层数据库的差异及数据位置的影响,降低应用和数据耦合,提高开放性和标准化,使大数据平台能够更高效地服务于上层应用与数据分析能力组件。数据透明访问模块将数据封装为独立的服务实体,提供标准化的交互接口,供内部模块或者外部系统调用,主要包括数据库访问路由、负载均衡、SQL 方言转换、数据适配、缓存加速、安全控制以及数据访问日志统计等功能。

实施数据透明访问主要实现的目标如下。

- 建立数据服务的标准化接口,降低数据依赖,促进

信息共享和应用重用;通过有效组织内部数据处理逻辑,提供标准化的对外服务接口,提高在安全、分布式事务处理方面的能力,降低应用与具体数据结构的依赖关系,提升应用的可扩展性和可重用性。

- 通过数据共享开放数据服务,提升与外部系统的互动能力;将数据共享为数据 Web 服务,提供同步响应、异步响应、发布订阅、数据库等多种数据服务模式,提高数据访问的开放性和灵活性。

数据透明访问模块是内部业务应用及外部系统与数据层进行数据交互的标准化接口,按功能又可分为数据访问接口、数据访问服务模块两部分,其架构如图 4 所示。

其中,数据访问接口通过消息模式、发布订阅模式、数据库模式等技术手段完成各种数据访问接口的适配。数据

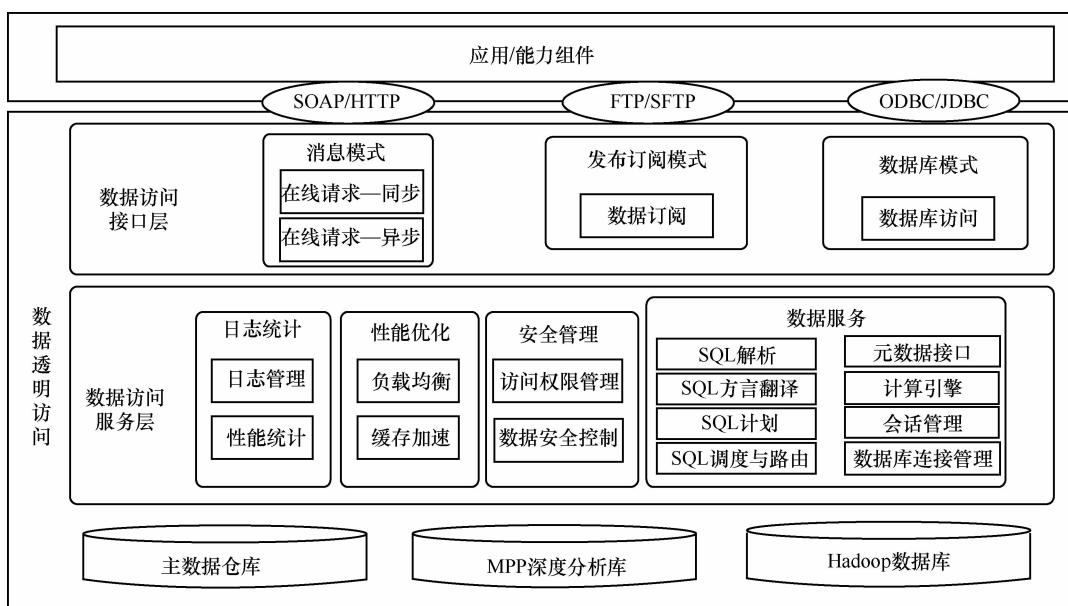


图 4 数据透明访问模块框架



访问服务层是具体的服务实体,完成具体的数据访问处理逻辑和数据操作等功能。数据访问服务主要负责对接收到的请求方的消息进行控制和管理,各类应用作为数据访问服务的请求方,发送请求消息给数据访问服务层,当数据访问服务模块接收到具体请求后,对请求的消息进行访问权限管理、SQL 解析、SQL 方言翻译,通过元数据查找数据具体位置,最后路由到相应数据库执行数据库操作,并返回结果给数据访问服务请求方。数据访问服务中的关键功能模块说明如下。

- 元数据接口:从外部导入元数据的接口,以了解数据、数据表在各个数据源的数据分布以及生命周期。
- SQL 解析:对 SQL 进行解析,拆解出表字段以及表关联关系,通过元数据获取表的物理位置。
- SQL 方言翻译:不同的数据库或者 SQL on Hadoop 引擎有不同的优化,可将通用的 SQL 转换成其专有的 SQL。
- SQL 计划:对 SQL 进行优化,转换成实际执行的 SQL。
- SQL 调度与路由:根据元数据优化对具体数据源的实际 SQL 执行。
- 计算引擎:归并各个数据源的返回结果。
- 数据库连接管理:管理实际连接各个数据库的连接。
- 会话管理:对所有访问作为会话进行管理控制,维护连接关系。
- 访问权限管理:对访问请求方进行用户权限控制,防止非法访问。
- 数据安全控制:对隐私数据和需要保护的数据根据用户访问权限进行加解密处理。
- 负载均衡:基于每个数据库服务器的访问负荷自动调节数据访问目的地,实现所有数据库服务器的负载分摊。
- 缓存加速:对热点访问数据进行缓存管理,后续直接访问缓存即可,既提高访问效率又减轻数据库负荷。

4.2 管理域

管理域包括数据治理和系统管理,核心是数据治理,包括元数据管理、数据质量管理、数据生命周期管理、数据安全治理。元数据管理对系统内所有数据进行定义与管理,是数据治理的核心,保证数据描述的准确性、一致性、端到端可管理。数据质量管理是确定哪种级别的数据质量属于“足够好”的质量,以满足上层应用的需求。数据生命周期管理就是制定数据存档策略,确保存储成本不会超出

控制,同时满足组织设定的数据保留计划,以便按照法规要求合理处置数据。数据安全治理是数据治理的另一个关键功能,企业必须严格关注并遵守用户隐私方面的法律法规,将用户真实标识与用户行为进行逻辑隔离,在架构上能够分开隐私域与非隐私域、敏感数据和非敏感数据、可对外呈现和不可对外呈现的数据。

大数据的数据量很大,配置管理元数据的工作量也会非常大。从方案上需要考虑从各数据源中直接获取元数据,要求支持手工获取和自动获取两种方式:对于自动获取方式,需要提供适配器管理,并能够提供 SQL 解析和 API 抽取元数据等功能;对于手工获取方式,需要提供与各种元数据相适应的元数据手工录入功能。

大数据平台需要端到端考虑数据的安全性,从物理安全、基础设施安全、网络安全、管理安全、数据安全 5 个层次都做好充分考虑。通过网络隔离,保证数据处理、存储安全和维护正常运行;通过对 Hadoop 集群内节点的操作系统进行安全加固等,保证节点正常运行;从身份认证、权限控制、审计控制等方面做好管理安全措施;从集群容灾、备份、数据完整性、数据保密性等方面,保证用户数据以及隐私数据的安全;用户隐私安全需重点关注,在数据架构上要考虑将用户真实标识与用户行为进行逻辑隔离,分开隐私域与非隐私域、敏感数据和非敏感数据、对外呈现可识别和不可识别等。

5 大数据建模关键技术

大数据平台能力层的核心是数据建模技术,大数据需要先进的建模挖掘技术才能发挥应有的价值,是实现互联网化存量经营的关键。模型的构建需要根据业务目标,合理选择目标函数 $y=f(x;\theta)$ 和优化函数参数 θ ,这个过程统称为数据建模。其中,目标函数的输入 x 和输出 y 需要业务人员的精确定义,而模型选择和具体参数调试或者长期的函数模型设计,需要数据挖掘和机器学习专家参与,因此大数据建模是一个需要跨领域合作的挑战性很高的项目。好的模型能输出业务人员期望看到的可执行知识,协助业务人员调整策略实现更好的存量经营。目前,互联网企业如百度、腾讯和阿里广泛使用大数据建模技术,以提供更佳的用户体验。本文简单介绍一下中国联合网络通信集团有限公司上海分公司(以下简称上海联通)采用若干大数据技术,通过用户离网预测大数据模型实现更优的存量经营,可以自动精准定位存量经营的用户对象目标,从

而以较小的成本实现较高的回报。

用户离网预测模型流程包括4个步骤:输入数据清理;特征提取和样本标注;随机森林模型训练;使用模型输出预测结果,如图5所示。

上海联通每个月预付费用户离网数量大约为20万户,如果挽留30%的用户,每个用户平均贡献75元左右,那么一年(12个月)的存量收入大约是5000多万元。但是,普通业务规则无法准确判断每个用户的离网倾向和离网的时间点,使得维系挽留的成本过高。例如,资源很大一部分投入离网倾向低的用户,并未提高挽留的用户总数。通过大数据建立模型可以更加精准地学习离网用户的画像,并针对性地设计维系挽留策略,从而使得有限的资源达到最优的使用效率。

用户离网预测模型中,输入的用户行为特征 x 是每个用户的历史话单、账单和网络侧记录。通过历史记录标注离网用户和在网用户,并通过两类用户的历史,例如提前1~2个月的记录,训练一个分类函数 $y=f(x;\theta)$ 。其中,目标 y 就是每个用户的离网概率或者倾向。通过历史标注的记录 $\{y,x\}$,利用机器学习算法估计模型的参数 θ 。模型参数确定好之后,可以利用模型函数,给定新的用户记录 x^{new} ,输出离网概率 $y^{new}=f(x^{new};\theta)$ 。考虑到用户离网数据随时间不断变化,需要动态地学习模型参数,使得预测精度有所保证。首先需要从话单、账单、客服数据和网络侧数据中抽取在网和离网用户的特征 x ,并根据业务人员定义的离网条件给出标注 y 训练模型。目前选择的是业界较先进的随机森林模型^[1],通过组合几百棵决策树,预测每个用户的离网倾

向。随机森林模型的目标函数如下:

$$P_{\text{离网}} = \sum_{b=1}^B p_b(x; \theta_b) \quad (1)$$

其中, x 是输入的用户电信行为特征, θ_b 是每一个决策树的参数, B 是随机森林中决策树的棵数,输出的离网概率 $p_{\text{离网}}$ 是 B 棵决策树输出的平均概率。需要优化随机森林参数 $\theta_b, 1 \leq b \leq B$,使得离网用户的概率高而在网用户的概率低。该模型的训练算法采用随机采样并对信息增益进行逐步排序挑选,实际速度很快,满足生产系统需求。如图5所示,模型输出每个用户在不同时间段的离网倾向,预测的离网用户名单被送往离网用户维系部门进行针对性的维系活动。

6 上海联通大数据应用成果

6.1 大数据平台及数据建模成果

上海联通已经完成大数据平台系统一期工程的建设,融合了B侧经营数据和O侧网络,逻辑组网如图6所示。

大数据建模的一个重要成果是使用模型输出的结果产生商业价值。以上海联通部署的大数据模型(用户离网预测模型)为例,介绍大数据建模的成果。

图7显示了区分离网和在网用户的主要特征并按照重要性(权重)进行排序。其中最重要的3个特征包括ARPU值(用户话费)、入网时长和通话时长,ARPU值即用户上个月的话费,是重要性最大的特征,其次是入网时长和通话时间,这3个特征的重要性占所有特征重要性的大约60%,其余52个特征仅占重要性的40%。这个结果表明,有一些关键特征可以将离网用户

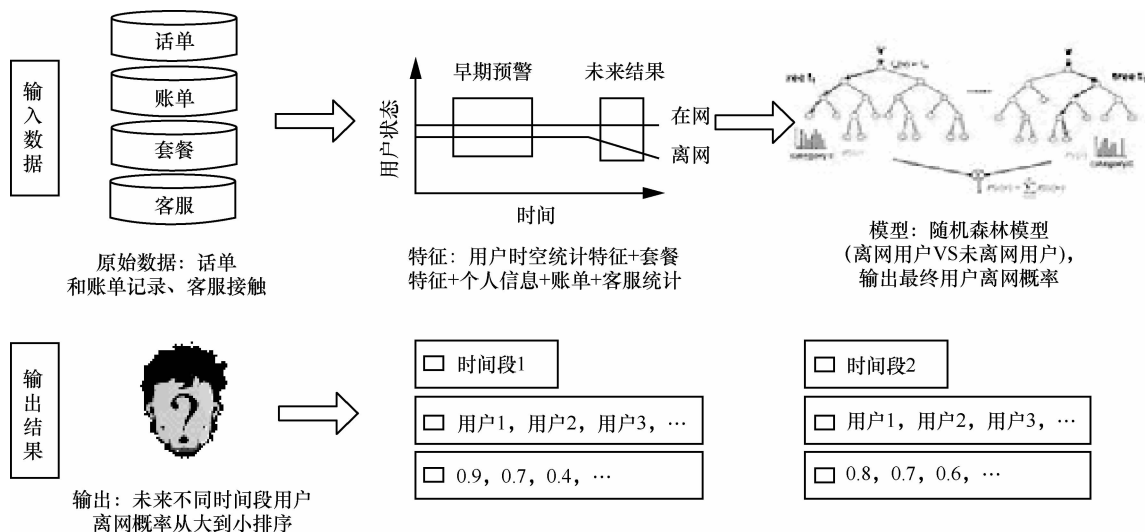


图5 用户离网预测模型流程示意

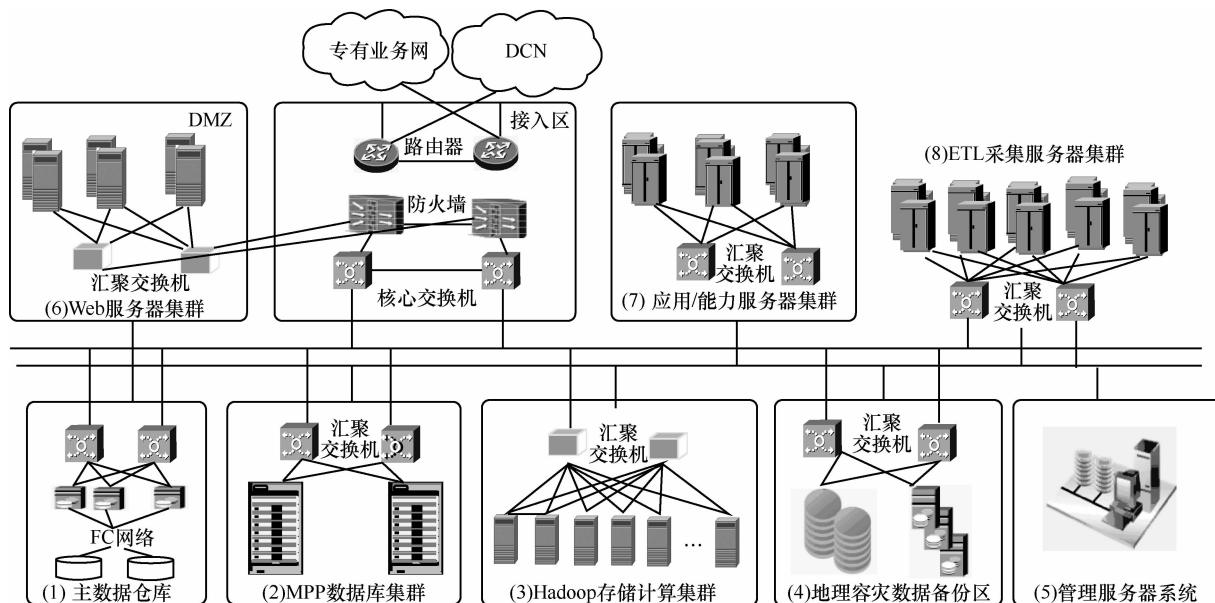


图6 上海联通大数据平台组网示意

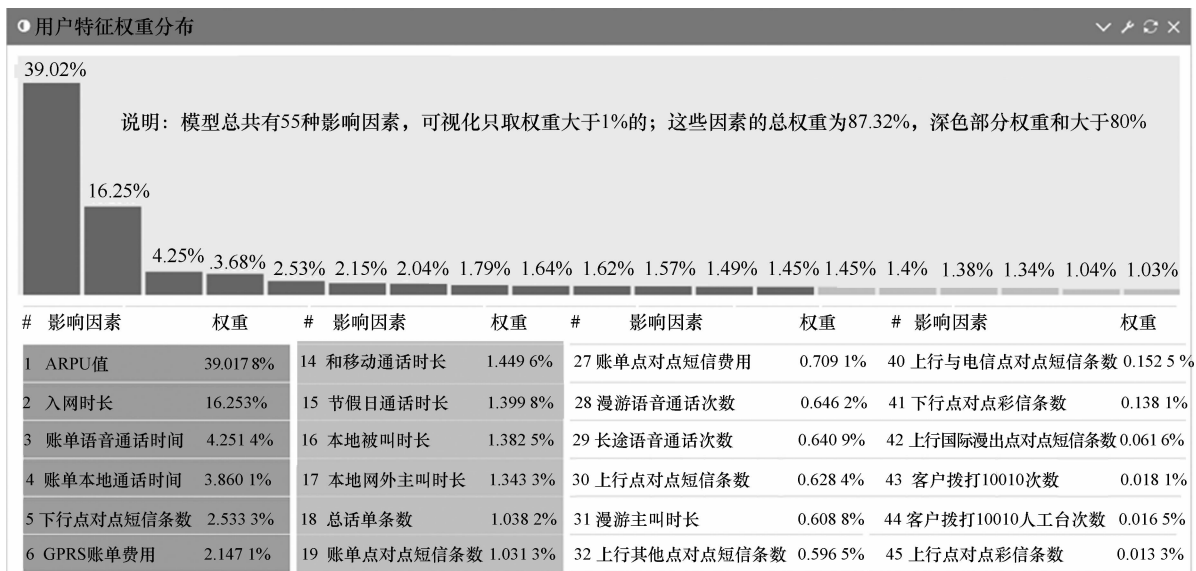


图7 用户离网预测模型输出不同特征的重要性

分辨出来，而模型挑选的特征也是与电信业务人员的经验相吻合的。

表1显示了预付费用户离网预测模型在实际应用中的表现。模型预测2014年5月份共467 820个用户有离网

表1 用户离网预测模型的实际效果

	在网用户	对照组	外呼组
总人数/人	2 379 806	457 543	10 277
充值人数/人	1 227 033	186 868	5 760
充值率	0.515 6	0.408 0	0.560 4

倾向,业务人员从中随机选择了10 277个用户外呼进行维系挽留活动(如充值送话费等活动),剩下的457 543个用户作为对照组不进行任何干预。业务人员评价模型的指标是“充值率”,期望看到的结果是对照组的充值率明显比在网用户的平均充值率低(即一部分用户离网),而外呼组比对照组的充值率高(即一部分用户被成功挽留)。表1的结果符合业务人员的预期。例如,对照组的充值率为0.408 0,低于在网用户平均充值率0.515 6大约20%;外呼组的充值率为0.560 4,相对对照组大幅提升约40%,

表明维系活动的确挽留了部分离网用户,提高了存量经营的效率。

6.2 存量经营商业效果

上海联通从 2013 年下半年开始启动存量经营大数据创新工作,至 2014 年上半年在用户维挽、通信价值提升、非通信价值变现等方面都取得了较好的成果。依托大数据平台,精细化流量相关业务的营销成功率较先前提升了 7.5 个百分点,客户对增值业务类产品和营销的满意度提升近 10 个百分点。首先,在用户维挽方面,累计公众存量用户的保有率同比提升 1.4 个百分点,累计后付费用户续约率同比提升 16 个百分点,续约后用户的 APRU 值有明显提升。其次,在通信价值提升方面,新增各类数据叠加包用户 59 万户,同比增长 190%,环比 2013 年第 4 季度增长 70% 左右。手机用户的数据业务收入占比同比提升 5.6 个百分点,3G 用户的户均流量同比提升 36%。最后,在非通信价值变现方面,与业界领先大数据挖掘公司签订基于大数据的互联网标签应用合作协议,与多家国际知名咨询公司签订基于移动轨迹信息的合作协议,在户外传媒广告价值评估、零售店面选址规划等项目中开展合作。

7 结束语

大数据技术应用于运营商存量经营是个崭新的课题。上海联通将存量经营和增量经营适当分离,利用大数据技术手段,真正关注和了解存量用户,持续挖掘和提升存量

客户的价值,是一种创新的尝试,为运营商在传统电信行业中经营转型做出有益探索。当前上海联通在大数据平台完成了一期建设,初步构建了统一的大数据采集与整合能力、分析处理能力以及应用能力。大数据建模方面已经开发完成多个数据模型,应用于多个商业场景,取得了较好效果。整体存量经营商业效果在 2014 年上半年取得明显成果。

从未来发展来看,大数据技术在运营商存量经营中的应用会越来越多。未来大数据会与云技术结合,具有更好的集群资源管理和处理能力。与物联网结合,获得更丰富的数据源信息。数据建模持续增强,例如在语音处理、视频处理、图分析、自然语言处理等方面实现突破,真正实现大数据可拟人视听的信息处理能力。数据可视化技术,会有更强的数据可视化展现能力、方便人类洞察数据的能力。

参考文献

- 1 Leo B. Random forests. *Machine Learning*, 2001, 45 (1): 5~32
- 2 Bishop C M. *Pattern Recognition and Machine Learning*. Springer, 2006
- 3 Brin S, Page L. The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems*, 1998(30): 107~117
- 4 Mhashikar K, Sarkar J. Formation integration: metadata management landscape. <http://www.docin.com/p-598461808.html>, 2006

(收稿日期:2014-06-01)

· 简讯 ·

我国未来网络试验设施首批规模试验应用启动及国际研讨会召开

2014 年 5 月,我国未来网络试验设施首批规模试验应用启动及国际研讨会在南京隆重举行。本次会议正式启动了我国未来网络试验设施首批规模试验应用,80 项未来网络前瞻性研究课题在该试验设施上开展试验验证,这是我国布局未来网络的又一重要举措,将有助于加快我国未来网络核心技术研发及产业化,形成示范效应,有助于国内相关科研力量的协同创新,推动我国未来网络试验设施项目实施,增强我国网络领域的国际竞争力,并为我国构建自主、可控、安全的网络环境及网络强国建设提供有力支撑。

本次研讨会正式启动了该试验设施的规模化试验应用,来自中国科学院、清华大学、北京邮电大学等科研院所的 80 项未来网络前瞻性研究项目获得了江苏省科技厅专项资金支持,在该设施上开展实验验证。通过规模化试验应用的开展,可以加快我国未来网络核心技术的研发及产业化进程,形成示范效应,有助于国内相关科研力量的协同创新,推动我国未来网络试验设施项目实施,增强我国网络领域的国际竞争力。

本次会议期间,江苏省未来网络创新研究院与中国香港、中国澳门、中国台湾等地科研机构及企业达成全面战略合作协议,将在未来网络技术研究、资源共享、产业化等方面展开合作,形成了“两岸三地”协同开展未来网络研发的格局,这必将有助于提升我国未来网络整体竞争力。