运营商面向大数据应用的数据脱敏方法探讨

乔宏明，梁奂

（中国电信股份有限公司广州研究院，广东 广州 510630）

【摘　要】　对当前数据脱敏的常用方法及其特点进行了阐释和分析，并结合运营商大数据应用的特点，从技术视角对各类脱敏方法的选择框架给出建议，供同业探讨。

【关键词】　数据脱敏　大数据　替换　混洗　数值变换

doi:10.3969/j.issn.1006-1010.2015.13.003　中图分类号：TP391.1　文献标识码：A　文章编号：1006-1010(2015)13-0017-04

引用格式：乔宏明,梁奂. 运营商面向大数据应用的数据脱敏方法探讨[J]. 移动通信, 2015,39(13): 17-20.

Discussion on Data Masking Oriented to Big Data Application for Operators

QIAO Hong-ming, LIANG Huan

(Guangzhou Research Institute of China Telecom Co., Ltd., Guangzhou 510630, China)

[Abstract]　Telecom operators in China focus on big data operation at present. The common methods and their features of data masking were addressed and analyzed in this paper. Combined with the features of operators' big data applications, the selection framework of different data masking methods was presented from technical perspective to provide a useful reference to peers.

[Key words]　data masking　big data　replace　shuffle　numerical conversion

1　引言

近两年来，国内通信运营商传统业务（特别是语音、短信）发展受互联网/移动互联网冲击的现象日益显现，业务量和业务收入增量的剪刀差越发明显：增长量差值（电信业务总量同比增长-电信业务收入同比增长）从2014年初的8%左右扩大到2015年一季度的近19%[1]，总体上“增量不增收”已经形成“惯性”。随着国家一系列导向性政策的出台（如《国务院办公厅关于加快高速宽带网络建设推进网络提速降费的指导意见》），当前运营商的“‘明星业务’——数据流量——收入预计”在未来1~2年内也将见顶，运营商现有的传统业务将逐步“公益化”。作为企业实体的运营商必须拓展新的业务领域，找到增长点，挖掘手中“大数据”这座金矿，成为三大运营商共同的目标。运营商数据分析应用范围在由“内”（经营分析、网络优化、针对性营销等）向“外”（广告、数据服务）转变，正在探索以“直接收益”为目标的大数据经营。

考虑到短期内运营商在新的大数据产业链中并不占据主导地位[2]，广泛的外部合作必不可少。但正是因为运营商所掌握的数据“含金量”过高，其使用也一直受到严格的政策监管：2013年7月19日，工信部下达了《电信和互联网用户个人信息保护规定》文件，明确了

收稿日期：2015-06-01

责任编辑：刘文竹　liuwenzhu@mbcom.cn

个人信息的收集和使用规范以及运营企业的管理责任。同时，当前广大客户对自身隐私的关注也日益增强，一直对运营商使用相关数据保持高度警觉状态，稍有疏忽就会给运营企业造成重大的声誉损失。运营商相关数据对外开放以进入大数据产业面临难以逾越的困难：既要挖掘使用，又要避免隐私风险。两难之间，对相关数据进行“脱敏”或许是一个解决方案。

2 常用数据脱敏架构和方法

所谓数据脱敏（Data Masking）是对个人身份识别数据（personal identifiable data）、个人敏感数据（personal sensitive data）和商业敏感数据（commercially sensitive data）进行伪装，以便用于生产系统以外的地方[3]。数据脱敏不是新的技术，当前也有很多成熟的商用解决方案可以选择，如Oracle的Data Masking组件[4]、IBM的InfoSphere Optim Data Privacy产品[5]、Informatica的Informatica Data Masking产品[6]等，其中Informatica的产品可以实现对异构数据的脱敏处理。针对特定的生产环境（如异构系统），也可以自己创建脱敏平台或系统进行脱敏处理。脱敏后数据的服务对象，可以是企业内部统计分析、企业生产系统的开放、测试环境，也可以是外部第三方。当然，面向不同的服务对象，针对其服务要求，脱敏的级别和方法也有不同。

从架构的角度看，数据脱敏有2种常用架构：

（1）动态（On the Fly/Dynamic）数据脱敏架构。指数据脱敏规则应用于在将数据从源数据库（生产库）导出到目标数据库（脱敏后数据库）的过程中进行脱敏处理，或者在生产系统产生实际数据的同时，也同步产生用于其他环境的脱敏数据。这种架构有2个好处：脱敏目标库可以获得实时性很高的数据；在生产系统外不存在非脱敏数据，减少安全风险。这种架构产生的问题是，脱敏处理会对生产系统产生一定的压力；脱敏策略可定制性不强，一旦投入持续生产就不能调整，否则会影响现有业务；脱敏应用会对源数据库到目标数据库链路安全和稳定性有较高要求；该架构一般都要求脱敏工具和生产库管理软件紧密耦合，限制可用工具的选择范围。

（2）静态（Static）数据脱敏架构。通过对源数据库的克隆来进行脱敏操作，形成目标数据库。脱敏规则可以在第三方实体上执行，也可以在目标数据库上执行。因为面对的是生产数据的镜像，这种架构可以根据需要调整脱敏规则，灵活性更高；脱敏工具的选择范围也更大；相对动态架构，静态架构对生产系统的压力更小。这种架构的风险是，因为涉及到第三方平台或目标数据库存储源数据，安全风险会增加；此架构获取的脱敏数据实时性相对动态架构偏低。

具体的数据脱敏方法，主要有以下6种：

（1）替代。指用伪装数据完全替换源数据中的敏感数据，一般替换用的数据都有不可逆性，以保证安全。替代是最常用的数据脱敏方法，具体操作上有常数替代（所有敏感数据都替换为唯一的常数值）、查表替代（从中间表中随机或按照特定算法选择数据进行替代）、参数化替代（以敏感数据作为输入，通过特定函数形成新的替代数据）等。具体选择的替代算法取决于效率、业务需求等因素间的平衡。替代方法能够彻底的脱敏单类数据，但往往也会使相关字段失去业务含义，对于查表替代而言，中间表的设计非常关键。

（2）混洗。主要通过对敏感数据进行跨行随机互换来打破其与本行其他数据的关联关系，从而实现脱敏。混洗可以在相当大范围内保证部分业务数据信息（如有效数据范围、数据统计特征等），使脱敏后数据看起来跟源数据更一致，与此同时也牺牲了一定的安全性。一般混洗方法用于大数据集合、且需要保留待脱敏数据特定特征的场景；对于小数据集，混洗形成的目标数据有可能通过其他信息被还原，在使用的时候需要特别慎重。

（3）数值变换。指对数值和日期类型的源数据，通过随机函数进行可控的调整（例如对于数值类型数据随机增减20%；对于日期数据，随机增减200天），以便在保持原始数据相关统计特征的同时，完成对具体数值的伪装。数值变化通过调整变动幅度可以有效控制目标数据的统计特征和真实度，是常用的脱敏方法。

（4）加密。指对待脱敏数据进行加密处理，使外部用户只看到无意义的加密后数据，同时在特定场

景下，可以提供解密能力，使具有密钥的相关方可以获得原数据。加密的方法存在一定的安全风险（密钥泄露或加密强度不够）；加密本身需要一定的计算能力，对于大数据集来源会产生很大资源开销；一般加密后数据与原始数据格式差异较大，“真实性”较差。一般情况下，加密的数据脱敏方式应用不多。

（5）遮挡（Mask Out）。指对敏感数据的部分内容用掩饰符号（如“X、*”）进行统一替换，从而使得敏感数据保持部分内容公开。这种方法可以在很大程度上脱敏的同时，保持原有数据感观，也是一种广泛使用的方法。

（6）空值插入/删除。指直接删除敏感数据或将其置为NULL值。在条件允许的情况下，这种方法最直接。

总体而言，数据脱敏的方法有以上6个类别。在具体应用时，可以根据业务需求，结合可用计算资源情况，进行灵活选择。

3 运营商大数据应用的特点

前期电信运营商在大数据应用方面主要聚焦在内部使用，隐私方面的风险相对可控，因而更多着力于管理流程和技术手段的完善，在脱敏方面投入的力量相对不大。当面向外部需要与第三方合作、甚至在政策范围内要输出部分数据给第三方时，数据脱敏就必不可少了。数据脱敏方法各有特色，具体选择需要结合运营商数据特点和实际的业务需求。

结合国内外各大运营商前期的探索研讨[7-8]，目前运营商面向外部的大数据应用主要包括以下场景：

◆精准广告：通信运营商发挥在用户上网行为数据采集方面的优势，为具有精准投放在线广告需求的企业客户筛选出高价值客户，提升其广告投放的精准性。

◆精准营销：通信运营商基于客户标签数据，为企业客户提供目标用户清单和广告精准推送方案，提升其营销活动效率。

需要说明的是，以上2种数据应用的开展目前也有政策限制。以短信这一最常用的广告营销手段为例，根据工信部最近发布的《通信短信息服务管理规定》[9]要求，从2015年6月30日起，短信息服务提供者、短信息内容提供者未经用户同意或者请求，不得向其发送商业性短信息。在此，必须假定运营商或外部第三方已经获得了向用户发送短信的许可。

◆数据报告：通信运营商基于通信流量数据挖掘结果，为行业客户提供流量、流向、应用活跃性等方面分析报告，为有相关需求的企业提供数据类咨询服务。

◆能力出租：通信运营商为不具备大数据运营能力的中小企业开放大数据平台的数据存储和分析能力，为其开展大数据应用提供高性价比的IaaS、PaaS平台。在运营商具备响应能力后，也可以进行“智力出租”，采用“来料加工”的方式，由客户提供数据，提出要求，运营商方面负责加工处理。

◆公共服务：对于政府或其他公共服务部门牵头开展的公共领域研究项目（如人群分布、人群流动、交通信息、舆情监控等），通信运营商作为重要的数据合作方和基础能力提供方，一方面提供部分通信相关数据，一方面出租IaaS/PaaS资源，实现企业和社会的双赢。

上述几个场景中，除能力出租外，都需要将电信数据和第三方数据紧密结合，才能获得预期成效，在此过程中数据的共享和开放不可避免。另一方面，这些大数据应用需要的往往都是用户敏感数据，涉及具体的客户信息必须做脱敏处理。

参照欧美运营商已经开展的大数据应用实践[10]，当前运营商的大数据应用有以下4个特点需要在脱敏方法选择时予以考虑：

（1）除国家特定要求外，输出数据不能包含个体性敏感信息。

（2）提供的数据应能提供较准确的统计性信息，支持群体性偏好或行为指标的深度分析。

（3）数据的时效性相对较高。特别是营销应用，大部分场景下都要求及时聚焦目标客户群，动态把握趋势和动向。

（4）数据需要能够动态更新：相对产业链上其他参与方，运营商最大的优势之一就是有源源不断的数据源，可以持续优化应用的效果（如客户刻画的精准度）。

4 脱敏方法选择框架

数据脱敏的最大难点在于平衡隐私保护和数据挖掘需求，从某种意义上，运营商必须要致力保护的隐私内容（具体某个用户的具体位置、社会关系、访问内容等）可能也正是外部第三方希望通过挖掘得到的内容。基于上述对运营商大数据应用特点的分析，结合具体应用场景，在选择脱敏方法时应该考虑以下6个因素：

（1）应用对数据可用性的要求，即脱敏后的数据满足分析应用需要的程度。如果脱敏后的数据完全无法用于目标分析，其也不具备使用价值。在特定的应用场景中，可能需要残留部分非关键信息（如手机号码部分字段、手机位置等）才能满足分析需求。

（2）应用对数据真实性的要求。这里的真实性是指脱敏后的数据对原有数据逻辑特征、统计分布特征的保留程度。绝大部分应用，特别是数据服务类应用对数据统计分布特征都有明确要求；同时对于复杂业务，其相关信息可能跨表跨库，数据间的逻辑特征也必须予以保留。

（3）应用对数据时效性的要求，即脱敏后数据需要在哪个时段内提供才有进一步分析挖掘的意义。

（4）应用对数据可重现性的要求，即相同参数配置下，相同源数据脱敏后的数据是否必须一致。

（5）脱敏方法资源占用。需要结合源数据量、源数据间行内同步、表内同步、跨表同步、跨库同步要求，考虑不同脱敏方法对计算资源、存储资源的需求。资源占用对数据时效性也会有潜在影响。

（6）脱敏方法可配置性。是否可以结合需求，通过对脱敏方法的配置生成个性化的脱敏后数据。

上述几个要素中，脱敏方法资源占用主要需考虑企业内部的资源约束，除此以外都和具体应用相关。针对需要数据输出的典型大数据应用，从业务需求视角，对现有的脱敏方法选择有以下简要分析供讨论，典型大数据应用如表1所示。

表1 典型大数据应用

应用类别	可用性	真实性	时效性	可重现	灵活性	适用方法
精准广告	高	高	中	高	高	遮挡、替代
数据报告	中	高	中	中	中	数值变换、混洗、删除、替代
精准营销	高	高	高	高	高	遮挡、替代
公共服务	中	中	高	中	中	替代、数值变换，混洗、删除

5 结束语

运营商开展大数据业务必须解决信息安全问题，为有效开展对外合作，必须对其持有的待挖掘数据进行脱敏。本文结合对数据脱敏的常用方法及其特点的理解，结合典型大数据应用类型，对各类脱敏方法的选择框架给出了建议。数据脱敏仅仅是运营商企业内部信息安全管理的一个环节，现有的脱敏方法既要服务于企业业务发展，也要遵从整体的IT安全治理要求，脱敏方案的制定和方法的选择需要业务需求单位（包括第三方）、IT安全监管单位和数据实际管控单位协同才能取得预期的成果。

参考文献：

[1] 中华人民共和国工业和信息化部. 2015年4月份通信业经济运行情况[EB/OL]. (2015-05-19)[2015-05-30]. http://www.miit.gov.cn/n11293472/n11293832/n11294132/n12858447/16594331.html.

[2] 乔宏明. 运营商在大数据产业中的定位刍议[J]. 移动通信, 2014(13): 16-17.

[3] WIKIPEDIA. Data masking [EB/OL]. (2015-04-23)[2015-05-30]. http://en.wikipedia.org/wiki/Data_masking.

[4] Oracle. Oracle Data Masking[EB/OL]. [2015-05-30]. http://www.oracle.com/technetwork/oem/app-quality-mgmt/default-1965435.html.

[5] IBM. InfoSphere Optim Data Privacy[EB/OL]. [2015-05-30]. http://www-03.ibm.com/software/products/en/infosphere-optim-data-privacy.

[6] Informatica. Data Masking[EB/OL]. [2015-05-30]. http://international.informatica.com/cn/products/data-masking/.

（下转第24页）

最大权重路径来进行切换路网标定，确定可以代表道路整体切换变化特征的“识别标签”，包括切换序列和相应的切换位置，后将其应用到手机用户的道路识别定位上，采用的是协同过滤算法的思路，很好地将该算法与现实难题相结合，较好地解决了利用低成本的通信切换数据进行交通状态跟踪的问题。并把该方法应用到实际的系统开发中，使保定市的道路人流量统计系统得以实现。

参考文献：

[1] 杨飞,裘炜毅. 基于手机定位的实时交通数据采集技术[J]. 城市交通, 2005(4): 63-65.

[2] 杨飞,惠英,杨东援. 基于手机切换定位的交通路网标定方法[J]. 同济大学学报：自然科学版, 2009,37(1): 67-72.

[3] 杨飞,惠英. 基于手机切换变化模式的道路匹配方法[J]. 系统工程, 2007,25(11): 6-13.

[4] 孙棣华,张星霞,张志良. 地图匹配技术极其在智能交通系统中的应用[J]. 计算机工程与应用, 2005(20): 225-228.

[5] 范秋明,何兆成. 基于手机基站定位数据的地图匹配研究[J]. 交通信息与安全, 2011(4): 52-57.

[6] 段玮. 基于协同过滤的个性化推荐算法研究[D]. 武汉: 华中科技大学, 2009.

[7] 郭艳红. 推荐系统的协同过滤算法与应用研究[D]. 大连: 大连理工大学, 2008.

[8] 邓爱林,朱扬勇,施伯乐. 基于项目评分预测的协同过滤推荐算法[J]. 软件学报, 2003,14(9): 1621-1626.

[9] 赵亮,胡乃静,张守志. 个性化推荐算法设计[J]. 计算机研究与发展, 2002,39(8): 986-991.

[10] 施华. 基于项目和用户双重聚类的协同过滤推荐算法[D]. 长春: 东北师范大学, 2009. ★

作者简介

杜翠凤：硕士毕业于广东外语外贸大学，现任广州杰赛科技股份有限公司研发中心大数据组项目经理，擅长于数据挖掘工具的应用，主要从事用户行为分析、用户轨迹分析。

余艺：大数据工程师，毕业于南京邮电大学软件工程专业，现任广州杰赛科技股份有限公司通信规划设计院研发中心大数据项目组研发工程师，擅长无线定位、用户感知分析、关联分析等方面的算法研究。

（上接第20页）

[7] 李曦烨. 大数据时代：通信运营五模式[EB/OL]. (2014-05-26)[2015-05-30]. http://labs.chinamobile.com/mblog/110109/221380.

[8] 黄小刚. 电信行业大数据应用的四个方向[J]. 信息通信技术, 2013(6): 26-28.

[9] 中华人民共和国工业和信息化部. 通信短信息服务管理规定[EB/OL]. (2015-05-28)[2015-05-30]. http://www.miit.gov.cn/n11293472/n11294912/n11296542/16613248.html.

[10] 36大数据. 全球十大电信巨头是如何玩大数据的[EB/OL]. (2014-04-19)[2015-05-30]. http:// www.199it.com/archives/210931.html. ★

作者简介

乔宏明：硕士毕业于北京邮电大学，现任职于中国电信股份有限公司广州研究院，主要研究方向为信息化规划和相关技术。

梁奂：国家正式注册咨询（投资）工程师，硕士毕业于暨南大学计算机软件专业，现任职于中国电信股份有限公司广州研究院，长期从事电信IT系统规划与IT服务管理方面研究工作。