中图分类号：TP309　文献标志码：B　文章编号：2095-641X(2018)01-0033-06　DOI：10.16543/j.2095-641x.electric.power.ict.2018.01.007
著录格式：王鑫，王电钢，母继元，等．基于机器学习的数据脱敏系统研究与设计 [J]. 电力信息与通信技术，2018,16(1)：33-38.

基于机器学习的数据脱敏系统研究与设计

王鑫，王电钢，母继元，常健，张凤
（国网四川省电力公司 信息通信分公司，四川 成都 610094）

Research and Implementation of Data Masking System Based on Machine Learning

WANG Xin, WANG Dian-gang, MU Ji-yuan, CHANG Jian, ZHANG Feng
(Information and Communication Branch, State Grid Sichuan Electric Power Company, Chengdu 610094, China)

摘要：近年来，国家电网公司各专业信息系统建设不断完善，为了保障数据在各类应用场景中安全使用，提出了一种基于机器学习的数据脱敏系统设计方案，重点针对敏感数据识别、定级，敏感算法制定，以及脱敏任务配置的实现方式进行研究。结合用户欠费信息的脱敏分析，验证了该方案具有自定义定级规则、辅助决策脱敏算法、配置脱敏任务等功能，为数据脱敏系统提供了一种智能化设计的新思路。

关键词：数据脱敏；机器学习；脱敏算法；静态脱敏；动态脱敏

Abstract: As various special information systems were built in State Grid Corporation of China, in order to ensure safely using data in all kinds of application scene, this paper proposes a data masking scheme based on machine learning, which focuses on researches on recognition and grading for sensitive data, data masking algorithm, and implementation of masking task configuration. Combining with actual application of user owe data masking analysis, it shows that this design has functions of custom grading rules, decision algorithm, and task configuration. Compared with traditional data masking method, this paper provides a new idea of intelligent design for data masking system.

Key words: data masking; machine learning; masking algorithm; static data masking; dynamic data masking

二等奖

0　引言

近年来，随着国家电网公司“三集五大”体系的推进，以及 SG186、SG-ERP 工程的建设，公司信息化实现了由分散到集中、由孤岛到共享的转变，积累了生产运行数据和经营管理数据约 5 PB，每月平均增长数据量约 46 TB，为数据集中共享和大数据分析、价值挖掘提供了有利条件 [1]。但是，数据资源中往往携带着有关用户与企业的敏感、隐私信息，一旦遭遇泄露、篡改，将给个人及公司甚至国家造成无法挽回的损失。因此，在数据共享使用过程中，如何准确定位敏感数据，合理制定脱敏策略，以达到数据安全可信、受控使用的目标，是一项亟待解决的技术问题。

数据安全问题的形势越来越严峻，数据脱敏逐渐受到企业的重视。传统的数据脱敏研究大多侧重于脱敏方法的实现 [2-4]，缺少权限判决、敏感识别等功能，系统化水平不够高。同时，脱敏算法的选择多为人工指定和自定义配置，智能化水平不够高。此外，模式识别的发展对实现脱敏信息的自动识别提

供了技术支持[5]，但在敏感信息分类定级问题上缺少对企业需求的考虑，专业化水平不高。

为解决数据脱敏的系统化、智能化、专业化水平不足等弱点，本文提出了一种独立于其他专业系统之外的数据脱敏系统。该系统同时集成了权限判决、数据分类、敏感信息识别、脱敏任务执行等功能；在敏感信息识别、敏感算法选择等关键环节采用文本分类、决策树等机器学习方法，可辅助人工实现脱敏策略制定；采用两层分类方式分类定级敏感信息，第一层按数据的专业和类型分类，第二层按规则进行分类定级。相较于传统数据脱敏方式，本文提供了一种智能化设计数据脱敏系统的新思路。

1 数据脱敏简介

数据脱敏又可称为数据去隐私化、数据变形，是指在保留数据初始特征的条件下，按需制定脱敏策略和任务，对敏感数据进行变换、修改的技术机制，可以在很大程度上解决敏感数据在非安全环境下使用的问题[6]。数据脱敏实现的难点在于如何同时保障数据的安全及其可用性，其关键就是脱敏算法的选择，就现阶段而言更多的是一种经验决策。根据不同的作用位置和实现原理，脱敏任务可分为静态脱敏（Static Data Masking，SDM）和动态脱敏（Dynamic Data Masking，DDM）。SDM一般用于非生产环境，在应用开发、测试、培训等场合中，为规避泄露风险，数据必须脱敏后才能被存储及使用。DDM常用于生产环境，当敏感数据被分析工具在线访问时，脱敏系统可以按照策略执行相应的脱敏算法。简言之，DDM与SDM的区别在于是否是在使用敏感数据时才进行脱敏。

数据脱敏系统应用框图如图1所示，本文构想了数据脱敏系统在国家电网公司的应用场景。用户或外部系统通过已集成的账号进入数据脱敏系统后，脱敏系统首先判断账号所具有的权限，并分配相应功能[7]。脱敏系统根据用户需求从各专业系统及公共系统抽取数据（包括结构化和非结构化数据），并对抽取的源数据进行分类、预处理、敏感识别定级以及选择脱敏算法和参数，完成脱敏策略制定。在变更脱敏任务时，用户可选择脱敏执行方式，其中静态脱敏可用于开发、测试以及数据迁移和存储；动态脱敏通过代理方式可为全业务统一数据中心等数据分析系统提供脱敏服务。如果没有新的数据或配置要求，脱敏策略和脱敏任务可以在脱敏系统中保存，以备后续调用及执行。

2 脱敏策略制定

从源系统抽取数据后，脱敏系统要为这些数据制定合适的脱敏策略。在策略制定阶段，系统需要着力解决敏感数据如何定级、是否需要脱敏、如何脱敏等一系列问题。

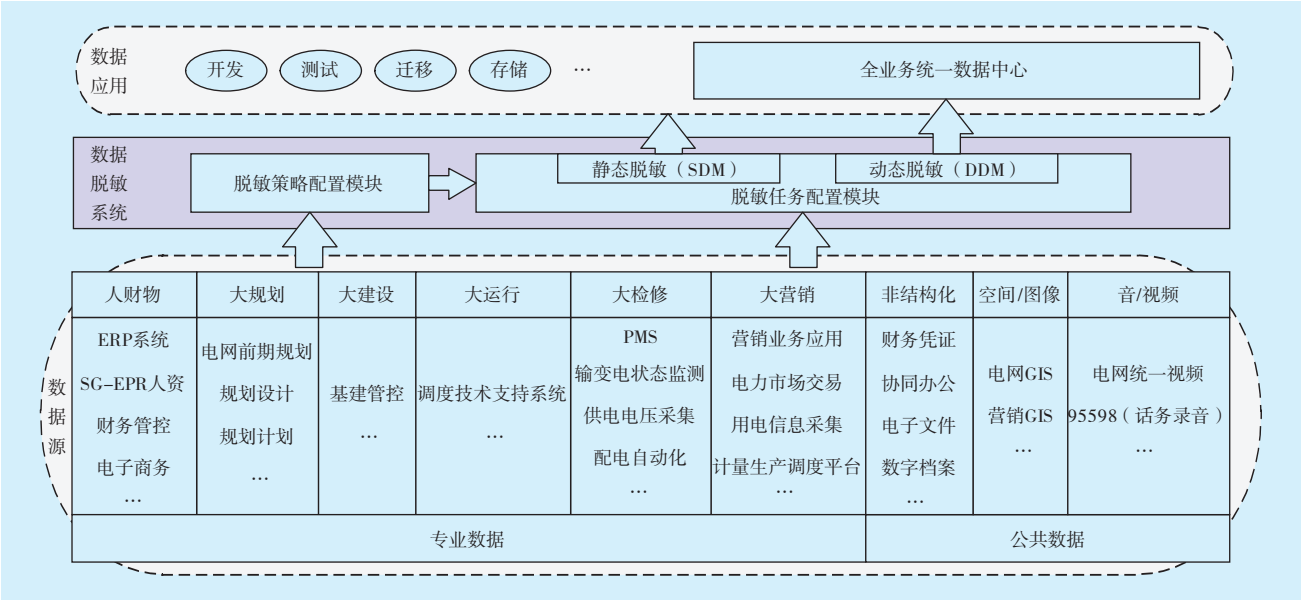

图1 数据脱敏系统应用框图

Fig.1 Application block diagram for data masking system

2.1 源数据分类及预处理

2.1.1 源数据分类

脱敏策略制定流程如图 2 所示。由于不同类型数据的敏感信息识别方法不同，系统需要对源数据分门别类。另外，同时识别多个专业的敏感信息也会为识别过程带来大量干扰，严重影响敏感信息识别的准确率[8]。根据文件格式类型，源数据可被分类为结构化数据、文本数据、图片、语音及视频数据。根据源业务系统不同，源数据可被分类为人财物、规划、建设、运行、检修及营销等数据。为了便于分类，本文系统分别为文本格式及业务系统分类设置了相应代码。

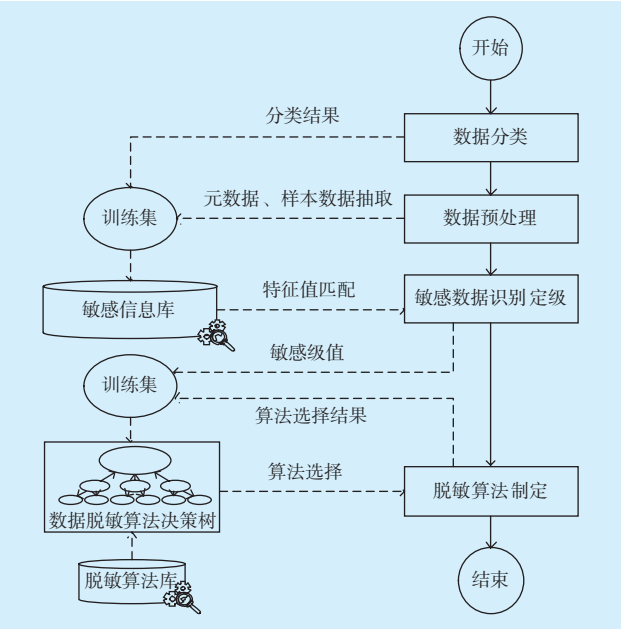

图 2 脱敏策略制定流程

Fig.2 Flow chart of masking strategy formulation

2.1.2 数据预处理

对源数据进行预处理以提取数据特征，通过数据特征匹配实现敏感信息识别[9]。脱敏系统采用自动化方式采集关系型数据库和非结构化系统的数据样本和元数据。结构化数据以数据字典（包括表名和字段名、类型、注释）的形式进行采集，并通过数据表遍历的方法从业务数据表中采集一定数量的样本数据。文本数据采用文本分词的方法对样本进行切割与合并，构建文本文件特征。对于图片、语音、视频数据，则通过相应领域的模式识别方法进行元数据和样本提取。元数据和样本采样完成后样本质量往往不佳，需要对其进行过滤和泛化处理，剔除数据“杂质”，以降低敏感信息识别与分类过程中的计算量[10]。

2.2 敏感数据识别定级

敏感数据识别是实现数据脱敏的关键前提。针对不同文件格式的数据，其敏感特征的检测方法会有所差异，数据脱敏系统应对其样本数据和元数据进行分类训练，最后分类建立敏感信息库。

敏感信息识别过程如图 3 所示，通过训练集获得文本、音频的语料库和图像视频的特征数据库，由安全部门和业务人员共同对语料库和特征数据库进行识别和分类[11]，选取其中具有代表意义的，可被标识为敏感信息的词、图像块、音频帧，形成敏感信息库，结合敏感信息模式匹配和源业务系统的重要程度，由人工辅助设定敏感级值，用于敏感信息定级。对预处理后的目标数据进行特征提取，将提取的特征值与敏感信息库的特征值进行匹配，当匹配命中时系统自动记录当前敏感信息的敏感级值。最后通过识别质量评估对错误分类进行纠正，并对未能识别的敏感信息进行补充。

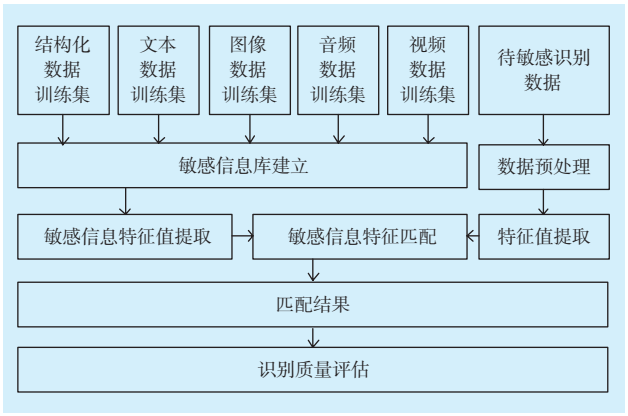

图 3 敏感信息识别过程

Fig.3 Sensitive information recognition process

2.3 脱敏策略制定

2.3.1 常用的脱敏方法

1）替换。替换（Replacement，RP）是指利用伪装数据对源数据中的敏感数据进行完全替换。为保证安全，一般替换用的数据都不具可逆性。

2）加密。加密（Encryption，EC）是指对待脱敏的数据进行加密处理，使外部用户或系统只能够接触无意义的加密数据。在特定场景下，系统可以提供解密能力，分发密钥给相关方以恢复原始数据。

3）遮掩。遮掩（Masking，MK）是指利用掩饰

二等奖

符号对敏感数据的部分内容进行统一替换，使得敏感数据保持部分内容公开。

4）删除。删除（Deletion，DL）是指直接删除敏感数据或将其置为空。

5）变换。变换（Change，CG）是指通过随机函数对数值和日期类型等源数据进行可控调整，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。

6）混洗。混洗（Shuffle，SF）主要是指通过对敏感数据采取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。

2.3.2 数据脱敏需考虑的因素

数据脱敏的最大难点在于平衡隐私保护和数据挖掘需求，脱敏算法适当与否直接影响到脱敏效果。为了制定合适的脱敏算法，结合具体应用场景，本文重点考虑了以下几个因素[12]。

1）可用性。即脱敏后的数据应能满足分析应用需求，若脱敏后的数据无法用于目标分析及应用，就不具备使用价值。在特定应用场景中，可能需要保留部分非关键信息（如身份证号码、手机号码的部分字段等）才能满足分析需求。

2）关联性。对于结构化和半结构化数据，在同一数据表中某字段与另外字段有对应关系，如果脱敏算法破坏了这种关系，该字段的使用价值将不复存在。通常在进行数据统计需要参考量的情况下，对数据的关联性要求较高。

3）真实性。脱敏后的数据对原始数据逻辑特征和统计分布特征的保留程度。为满足这种特性，数据的原始值需要尽可能地被保留。

4）时效性。数据提供需要有一定的及时性，超过一定时间后脱敏数据可能就不再具有进一步分析挖掘的意义。因此，应尽量避免使用耗时的脱敏算法，比如加密算法。

5）可重现。即相同源数据在配置相同算法和参数的情况下，脱敏后的数据应保持一致，随机类的算法应避免使用。

6）可配置。主要是指可以灵活配置、组合脱敏算法，可以结合不同需求生成个性化的脱敏数据。

由于上述各因素需要付诸实际应用才有意义，脱敏算法与脱敏效果之间的关系只能作定性分析。决策树是一种简单而又被广泛使用的分类器，具有描述性，有助于人工分析，同时决策树只需一次构建，可反复使用[13]。对敏感级值和6个因素进行量化，以具有代表性的应用场景来构建选择脱敏算法所需的训练集，形成决策树。利用决策树可以高效地对脱敏数据进行算法推荐，辅助系统用户进行算法选择。新的脱敏应用发生后，其敏感级值和算法选择结果将加入训练集，逐步对决策树进行完善，从而提高决策树的鲁棒性。

3 脱敏任务配置

在完成脱敏策略制定后，为使脱敏任务能够长时间工作，首先需在脱敏系统中对源数据所在业务系统的地址及端口号进行注册[14]。然后，获取已制定的脱敏策略，脱敏系统按照选择的脱敏算法及相关参数生成脱敏代码。用户根据应用场景选择脱敏实现方式，对于静态脱敏，系统先执行脱敏操作，并将脱敏结果缓存在本地存储，待目标系统需要获取脱敏数据时，用户在脱敏系统中注册目标系统的地址及端口，最后将本地脱敏数据传输至目标系统。而对于动态脱敏，用户必须先在脱敏系统中注册目标系统地址、端口以及目标系统使用账号，然后将脱敏代码下发至代理服务器，由代理服务器进行在线数据脱敏，并将脱敏结果返回至脱敏系统，最后传输至目标系统，由目标系统中的数据需求方的账号使用。脱敏任务配置流程如图4所示。

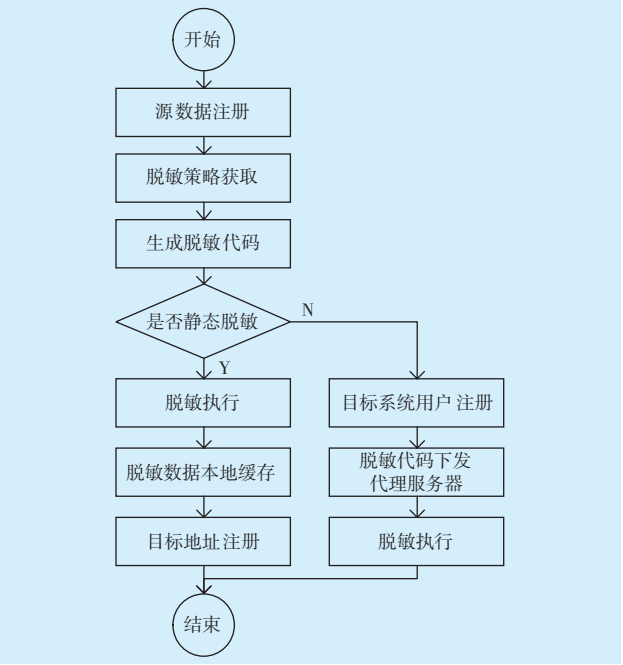

图4 脱敏任务配置流程

Fig.4 Configuration flow chart of data masking task

二等奖

4 应用验证

本文以营销业务应用系统数据库中的实收电费信息表为例[15]，分析人员希望通过分析某一用电用户的欠费历史，对该用户的用电行为进行评估。实收电费信息表中涉及的用户用电欠费信息和缴费信息均为敏感信息，因此在使用数据表进行分析之前，应对其进行脱敏处理。

在明确敏感信息后，脱敏实现的关键点为脱敏算法的选择。脱敏算法选择决策树如图 5 所示，为数据脱敏系统中用于脱敏算法制定的决策树，由 192 组数据的训练集训练得到，其中 Level 表示敏感级值，ZSX、KYX、KPZ、GLX、SXX、KCX 分别为真实性、可用性、可配置、关联性、时效性、可重现等 6 个脱敏算法选择因素的中文首字母缩写，RP、EC、SF、DL、MK、CG 为上文介绍的 6 种脱敏方法。在利用决策树选择脱敏算法之前，脱敏系统用户应与分析人员共同确定敏感信息在本次分析过程中的 6 个因素是否满足。

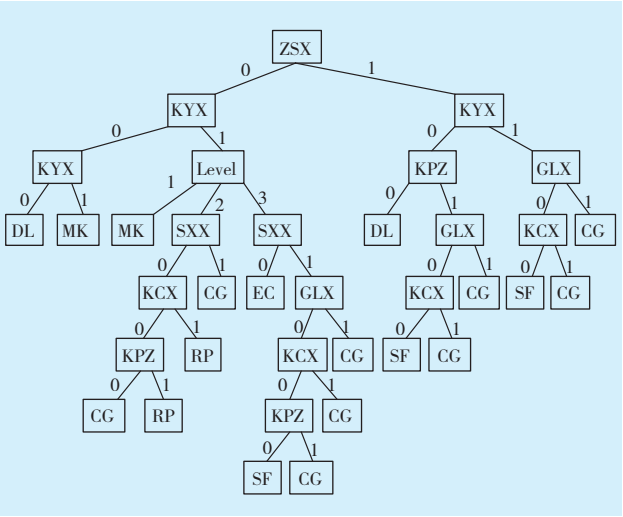

图 5 脱敏算法选择决策树

Fig.5 Decision tree for masking algorithm selection

欠费信息是统计分析的对象，数据应具有可用性和真实性。同时，由于是分析某一用户，因此用户编号字段和欠费信息间的对应关系不能被破坏，数据需要具有关联性，而时效性、可重现和可配置 3 个因素在这里则无关紧要。因此，综合考虑因素分析，系统用户可按照图 5 决策树选择数据变换（CG）算法对欠费信息进行脱敏处理。而缴费信息不是统计分析的对象，不具可用性，同时另外 5 个因素也不用具备，根据决策树推荐可以直接删除并置空缴费信息。

在系统用户完成静态脱敏任务配置后，脱敏系统可按照选择的脱敏策略对实收电费信息表进行脱敏，待分析人员的目标地址及端口在脱敏系统中注册后，即可将脱敏表从本地缓存传给用于分析的目标系统。

5 结语

数据脱敏将成为大数据时代企业数据化运行维护的必要安全机制。本文从国家电网公司信息化建设实际情况出发，分析并提出了一种数据脱敏系统设计方法，探讨利用机器学习方法使脱敏过程更加系统化、智能化、专业化，并结合实际应用场景验证了脱敏策略制定功能。随着未来对数据脱敏的进一步研究，脱敏系统将可实现更细粒度的访问控制、更精确的需求理解能力、更强的扩展能力以及更友好的交互方式，从而满足更多跨系统、跨专业、跨行业的数据交互、共享和融合需求。

二等奖

参考文献：

[1] 王玮, 刘荫, 于展鹏, 等. 电力大数据环境下大数据中心架构体系设计[J]. 电力信息与通信技术, 2016, 14(1): 1-6.
WANG Wei, LIU Yin, YU Zhan-peng, et al. System design of the big data center architecture in electric power big data environment[J]. Electric Power Information and Communication Technology, 2016, 14(1): 1-6.

[2] SWEENEY L. K-anonymity: a model for protecting privacy[J]. International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, 2002, 10(5): 557-570.

[3] 王茂华, 郝云力, 褚亚伟. 具有隐私保护功能的数据加密算法[J]. 计算机工程与应用, 2014, 50(23): 87-90.
WANG Mao-hua, HAO Yun-li, CHU Ya-wei. Data encryption method based on privacy preserving[J]. CEA, 2014, 50(23): 87-90.

[4] 周期律, 张旭春, 蔡仕志. 商业银行客户姓名脱敏技术研究[J]. 中国金融电脑, 2014(6): 65-68.
ZHOU Qi-lv, ZHANG Xu-chun, CAI Shi-zhi. The research of customer name desensitization technology of commercial banks[J]. Financial Computer of China, 2014(6): 65-68.

[5] 崔星华. 基于局部特征的图像模式识别算法研究[J]. 吉林建筑工程学院学报, 2014, 31(6): 52-54.
CUI Xing-hua. Research on image pattern recognition algorithm

based on local feature[J]. Journal of Jilin Institute of Architecture & Civil Engineering, 2014, 31(6): 52-54.

[6] 陈天莹, 陈剑锋. 大数据环境下的智能数据脱敏系统[J]. 通信技术, 2016, 49(7): 915-922.
CHEN Tian-ying, CHEN Jian-feng. Intelligent data masking system for big data productive environment[J]. Communications Technology, 2016, 49(7): 915-922.

[7] 谭晶, 仇红剑, 徐明生, 等. 电力企业大数据平台数据权限控制机制研究与应用[J]. 电力信息与通信技术, 2017, 15(5): 49-53.
TAN Jing, QIU Hong-jian, XU Ming-sheng, et al. Research and application of data access control in big data platform of power enterprise[J]. Electric Power Information and Communication Technology, 2017, 15(5): 49-53.

[8] 刘金. 基于数据特征的敏感数据识别方法[J]. 信息通信, 2016(2): 240-241.
LIU Jin. Sensitive data recognition method based on data features[J]. Information & Communications, 2016(2): 240-241.

[9] 王丹, 李建岐, 廖斌. 基于优化去雾算法的配网开关状态视频识别技术研究[J]. 电力信息与通信技术, 2017, 15(10): 31-37.
WANG Dan, LI Jian-qi, LIAO Bin. Research on video recognition technology of distribution switch based on optimized defogging algorithm[J]. Electric Power Information and Communication Technology, 2017, 15(10): 31-37.

[10] 凌笑, 易衍孜. 基于电网GIS平台的电网资源图形数据质检工具[J]. 电力信息与通信技术, 2017, 15(2): 23-26.
LING Xiao, YI Yan-zi. Graphical data quality check toolset for power grid resource based on power grid GIS platform[J]. Electric Power Information and Communication Technology, 2017, 15(2): 23-26.

[11] 吕辉, 许道强, 仲春林, 等. 基于电力大数据的标签画像技术与应用研究[J]. 电力信息与通信技术, 2017, 15(2): 43-48.
LV Hui, XU Dao-qiang, ZHONG Chun-lin, et al. Study on tag portrait technology based on electric power big data and its application[J]. Electric Power Information and Communication Technology, 2017, 15(2): 43-48.

[12] 苏亮, 陈亚军. 基于ID3决策树算法在环境监测的应用[J]. 信息通信, 2015(6): 22-23.
SU Liang, CHEN Ya-jun. Application of ID3 decision tree algorithm in environmental monitoring[J]. Information & Communications, 2015(6): 22-23.

[13] 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, 24.
QIAO Hong-ming, LIANG Huan. Discussion on data masking oriented to big data application for operators[J]. Mobile Communications, 2015, 39(13): 17-20, 24.

[14] 郝赫, 胡学勇, 王国娟, 等. 国家电网公司典型系统一级部署模式集成方案[J]. 电力信息与通信技术, 2014, 12(7): 69-73.
HAO He, HU Xue-yong, WANG Guo-juan, et al. Typical system integration solution with one-level deployment mode in state grid corporation of China[J]. Electric Power Information and Communication Technology, 2014, 12(7): 69-73.

[15] 黄文思, 郝悍勇, 李金湖, 等. 基于决策树算法的电力客户欠费风险预测[J]. 电力信息与通信技术, 2016, 14(1): 19-22.
HUANG Wen-si, HAO Han-yong, LI Jin-hu, et al. Prediction of power customer arrear risk based on decision tree algorithm[J]. Electric Power Information and Communication Technology, 2016, 14(1): 19-22.

编辑　邹海彬

收稿日期：2017–11–20

王鑫

作者简介：

王鑫（1990–），男，助理工程师，从事信息运维工作，491543245@qq.com；

王电钢（1973–），男，高级工程师，从事信息运维管理工作；

母继元（1978–），男，高级工程师，从事信息运维工作；

常健（1979–），男，工程师，从事信息运维工作；

张凤（1986–），女，工程师，从事信息运维工作。

二等奖